OxWaSP: Probability & Approximation

Lecture 1: classical probabilistic convergence

Wilfrid S. Kendall[1]

Department of Statistics, University of Warwick

22 October 2018

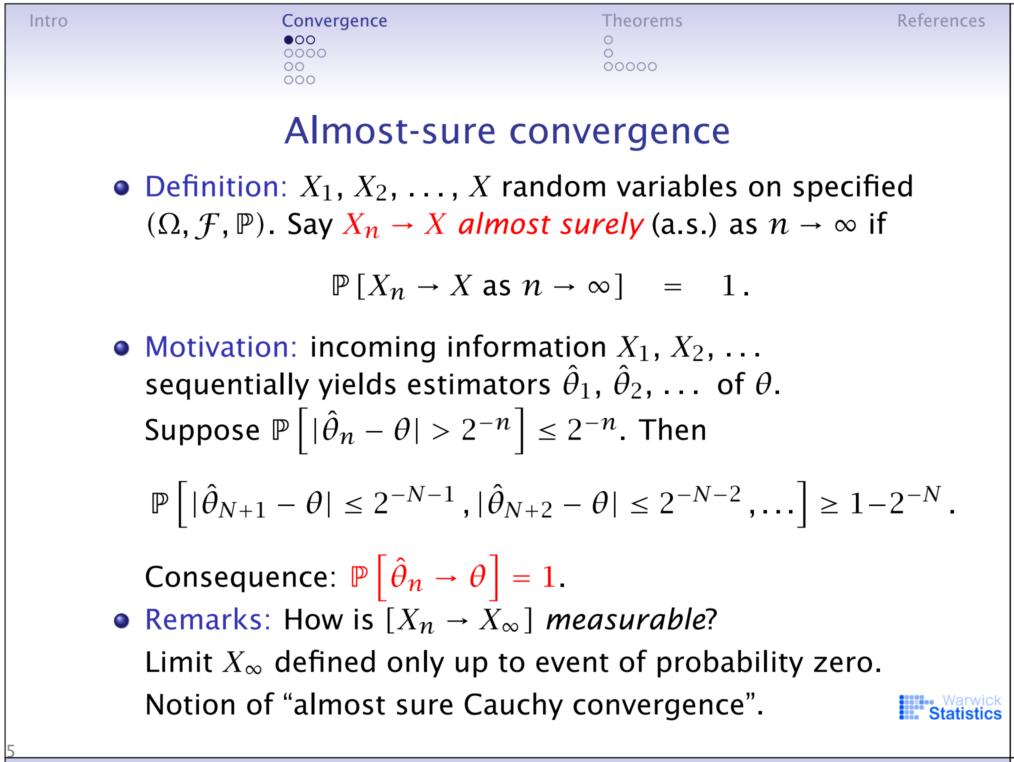
[1] w.s.kendall@warwick.ac.uk

Dilemma:
All of you are very able students, but with very diverse backgrounds.
Some OxWaSP students know lots of probability.
Some OxWaSP students only know a little.
This module seeks to compromise: providing overview of basics and introducing some particular topics that even the most probabilistic may not yet have encountered.

Introduction

- Objective: To expose students to the study of probabilistic approximation, important in various areas of probability and statistical applications.
- This lecture:
  - Convergence as an idealized form of approximation;
  - Review classical convergence theory, with examples;
  - Review classical theorems.

This morning session: 50 minutes of lecture, 25 minutes to try out some exercises, join together for 15 minutes recap and (perhaps) discussion of some related project ideas.



Almost-sure convergence

- Definition: $X_1, X_2, \ldots, X$ random variables on specified $(\Omega, \mathcal{F}, \mathbb{P})$. Say $X_n \to X$ *almost surely* (a.s.) as $n \to \infty$ if

$$\mathbb{P}\left[X_n \to X \text{ as } n \to \infty\right] \quad = \quad 1\,.$$

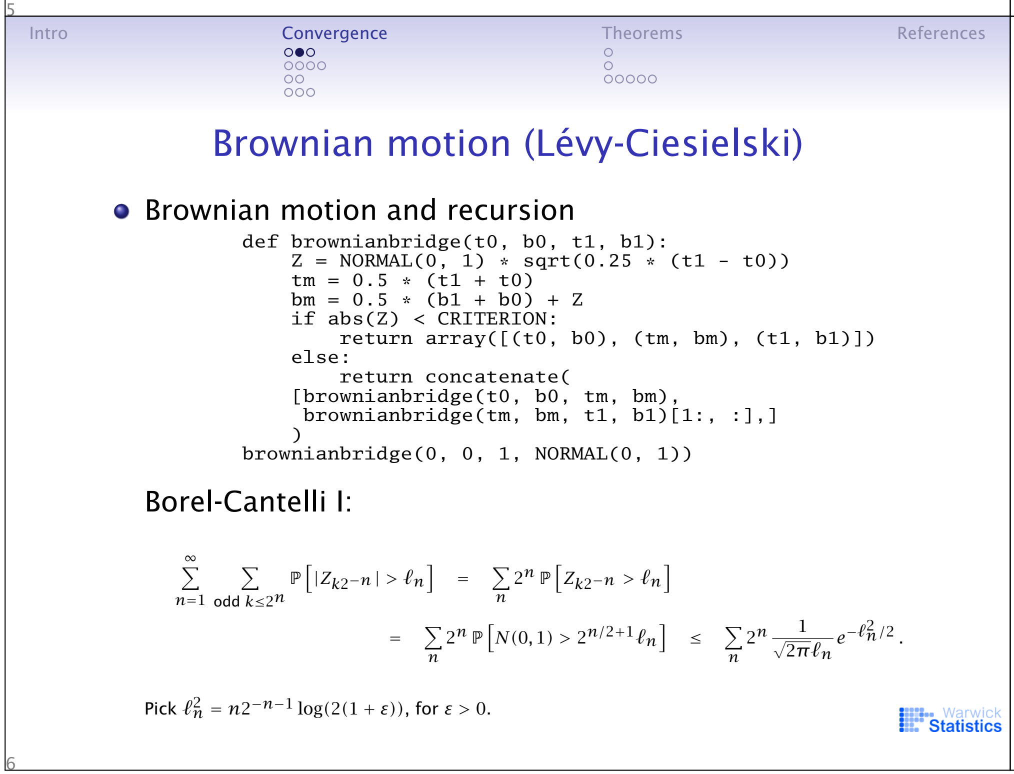
- Motivation: incoming information $X_1, X_2, \ldots$ sequentially yields estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$ of θ. Suppose $\mathbb{P}\left[|\hat{\theta}_n - \theta| > 2^{-n}\right] \leq 2^{-n}$. Then

$$\mathbb{P}\left[|\hat{\theta}_{N+1} - \theta| \leq 2^{-N-1}, |\hat{\theta}_{N+2} - \theta| \leq 2^{-N-2}, \ldots\right] \geq 1 - 2^{-N}\,.$$

  Consequence: $\mathbb{P}\left[\hat{\theta}_n \to \theta\right] = 1$.
- Remarks: How is $[X_n \to X_\infty]$ *measurable*?
  Limit X_∞ defined only up to event of probability zero.
  Notion of "almost sure Cauchy convergence".

We begin by reviewing various kinds of convergence typically encountered in undergraduate probability. This material will be well-known to some, but others will not have encountered this systematically. Useful references: Grimmett and Stirzaker (2001) is a good basic reference; see also the much more advanced treatment by Chow and Teicher (2003).

1. The definition of almost-sure convergence is *conceptually simple* (it "probabilizes" a concept from classical analysis).
2. Use: sub-additivity of probability. Explain relevance of Borel-Cantelli I.
3. The definition of almost-sure convergence is *structurally complicated* because it involves the whole sequence in a single probability statement.
4. Review Borel-Cantelli lemma(s) in light of above. Note: BC2 holds under only pairwise independence!
5. Exercise 1.1 (1): express event $[X_n \to X]$ in terms of more basic events; Exercise 1.1 (2): Cauchy convergence.



Brownian motion (Lévy-Ciesielski)

- Brownian motion and recursion

```
def brownianbridge(t0, b0, t1, b1):
    Z = NORMAL(0, 1) * sqrt(0.25 * (t1 - t0))
    tm = 0.5 * (t1 + t0)
    bm = 0.5 * (b1 + b0) + Z
    if abs(Z) < CRITERION:
        return array([(t0, b0), (tm, bm), (t1, b1)])
    else:
        return concatenate(
    [brownianbridge(t0, b0, tm, bm),
     brownianbridge(tm, bm, t1, b1)[1:, :],]
    )
brownianbridge(0, 0, 1, NORMAL(0, 1))
```

Borel-Cantelli I:

$$\sum_{n=1}^{\infty} \sum_{\text{odd } k \leq 2^n} \mathbb{P}\left[|Z_{k2-n}| > \ell_n\right] \quad = \quad \sum_n 2^n\, \mathbb{P}\left[Z_{k2-n} > \ell_n\right]$$

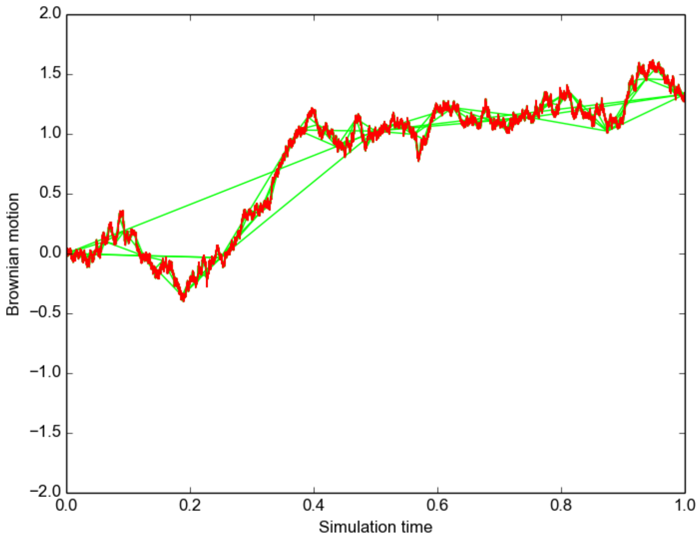
$$= \quad \sum_n 2^n\, \mathbb{P}\left[N(0,1) > 2^{n/2+1}\ell_n\right] \quad \leq \quad \sum_n 2^n \frac{1}{\sqrt{2\pi}\ell_n} e^{-\ell_n^2/2}\,.$$

Pick $\ell_n^2 = n2^{-n-1}\log(2(1+\varepsilon))$, for $\varepsilon > 0$.

HISTORY: Brown, Einstein, Wiener, Ciesielski, Lévy, ...
Fascinating for probabilists in late C20,
a crucial tool for probabilists and statisticians in early C21.
Basic model for random continuous variation in time, often arising (lightly modified) as limiting approximation.

1. Elegant recursive construction, related to Haar / Schauder wavelets.
2. Various `numpy` functions are used here. In particular, NORMAL derived from (for example) `numpy.random.normal`, but take care of standard deviation *versus* variance!
3. If the probabilities produce a finite sum, then (Borel-Cantelli I) only finitely many of the events can occur.
   We deduce this using the Mills' ratio inequality, and a strategic choice of level ℓ_n,
4. Mills' ratio $\overline{\Phi}(x)/\phi(x)$ for $x > 0$, where ϕ is the standard normal density and $\overline{\Phi}$ is the complementary distribution function.
   Use $\overline{\Phi}(x) = \int_x^\infty \phi(u)\, \mathrm{d}u \leq \int_x^\infty (u/x)\phi(u)\, \mathrm{d}u = \phi(x)/x$.

Brownian motion (Lévy-Ciesielski) illustrated



Various iterations of the Lévy-Ciesielski recursive construction, using different seeds for each image.

1. Each stage of the iteration is piecewise-linear, continuous (but slopes tending to increase as iteration progresses).
2. Almost surely, (uniform) convergence.
   Hence continuous limit.
3. Almost surely *nowhere* differentiable!
4. There is an immense literature on the theory of Brownian motion!

Convergence in probability

- Definition: $X_1, X_2, \ldots, X$ on specified $(\Omega, \mathcal{F}, \mathbb{P})$. Say *$X_n \to X$ in probability* (in prob.) as $n \to \infty$ if, for all $\varepsilon > 0$,
$$\mathbb{P}\left[|X_n - X| > \varepsilon\right] \quad \to \quad 0\,.$$
- Motivation: Suppose only $\mathbb{P}\left[|\hat{\theta}_n - \theta| > 2^{-n}\right] \leq \frac{1}{n}$.

  Then $\mathbb{P}\left[\hat{\theta}_n \to \theta\right]$ can be zero!

  Convergence in probability includes threshold cases when almost sure convergence does not quite happen.
- Remarks: Convergence in probability only involves two random variables at a time (θ_n and θ).

  $X_n \to X$ *a.s.* implies $X_n \to X$ *in prob.*

  "Fast" $X_n \to X$ *in prob. does* imply $X_n \to X$ *a.s.*

The relationship between convergence almost surely and convergence in probability is quite subtle, but is worth getting straight.
It is to do with the difference between (a) a sequence which will eventually draw arbitrarily close to its limit,
and (b) a sequence which in due course is very likely to be close to the limit, but may occasionally exhibit large fluctuations.

1. The definition of convergence in probability is conceptually more awkward albeit structurally more simple.
2. The sub-additivity argument experiences difficulties (because $\sum \frac{1}{n} = \infty$, so the lower bound in becomes negative and therefore useless). Sometimes we can fix things up by replacing the 2^{-n} in $|\hat{\theta}_n - \theta| > 2^{-n}$ by some larger quantity, and reducing the sequence of $\frac{1}{n}$ to some finitely summable sequence.
   Sometimes we can't. (For example: in case of independence, consider $X_n = 1$ with probability $1/n$, $X_n = 0$ otherwise, use Borel-Cantelli II.)
3. The parameter θ is viewed as random in the Bayesian world, because it is unknown! Even in the classical world, it is random albeit in a degenerate sense.
4. Exercise 1.2: (1) almost sure convergence implies convergence in probability (here it really pays to elaborate on $[X_n \to X]$ in terms of countable unions and intersections!);
   (2) counterexample to possible converse;
   (3) "fast" convergence in probability (or suitable subsequence) forces almost sure convergence.

Empty holes in Poisson histograms (I)

- Example (WSK 1993).
- Observe $X_1, X_2, \ldots$, conditionally independent Poisson(μ), prior parameter μ.
- Estimate μ from observation $X_{N+1} = x$ under quadratic loss, using previous $X_1, \ldots \; X_N$ for information about distribution of μ.
- Bayes estimate $\mathbb{E}\left[\mu \mid X_{N+1} = x, X_1, \ldots, X_N\right]$; replace distribution of μ by previous empirical distribution.
- (Empirical Bayes) estimator:

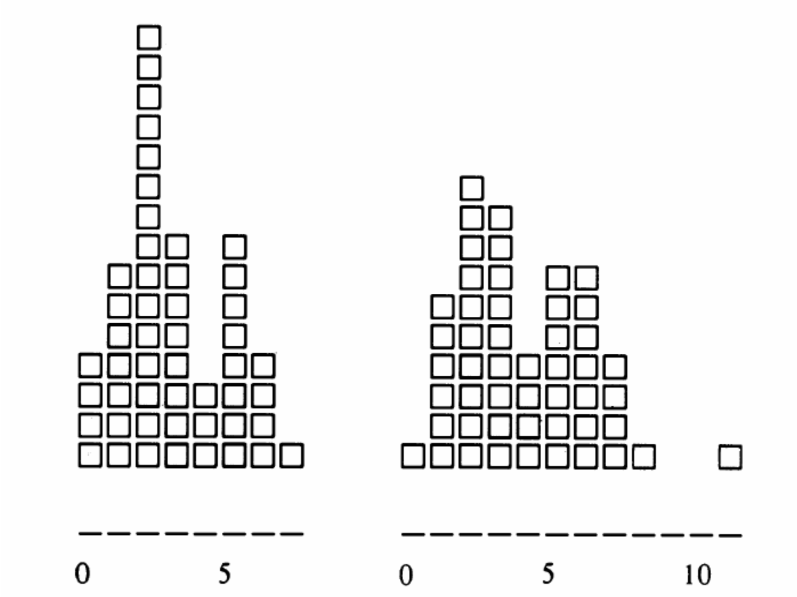
$$\frac{(x+1)\,\#\{r \leq N : X_r = x+1\}}{1 + \#\{r \leq N : X_r = x\}}\,.$$

- Not good (too small) if no values of $x+1$, bad if no values of x.
- Can there be empty holes?

This arose from moderation of a Warwick Statistics examination question.

1. “Empty holes exist”: histogram is disconnected.
2. This example is a graphic demonstration of how sequences convergent only in probability will (as one runs down the sequence) occasionally exhibit large deviations from the limit.
3. Less naïve empirical Bayes estimators use, for example, smoothing to ameliorate problems.
4. Simulation is indecisive: perhaps yes, perhaps no.

Empty holes in Poisson histograms (II)



Left-hand: one component.
Right-hand: two components seprated by two empty cells.

Empty holes in Poisson histograms (III)

- Fix prior parameter μ.
- Study C_N, number of components of the histogram of $X_1, \ldots \; X_N$, as N increases.
- Computations show: $C_N \to 1$ in probability as $N \to \infty$
- However C_N does not converge almost surely (Borel-Cantelli argument).
- Failure of almost sure convergence occurs because infinitely often (but with increasing rarity) we see single isolated cells containing one value of X at the right of the histogram, separated from the main body by a single empty cell.
- Occasionally the histogram will have been disconnected for arbitrarily large proportions of the past.

Note that this is a simple example of a generic phenomenon in mathematical science: simulation experiments may be indecisive on their own because the event under consideration (here, failure of connectedness of the histogram) becomes increasingly rare as the size N of the experiment increases, but nevertheless will keep on recurring as N increases. Simulation is not always enough!

Suppose μ is random with a prior distribution. We could consider a conjugate prior, namely some Gamma distribution. Then infinitely often we will see specified finite patterns of empty cells at the right-hand edge. So naïve empirical Bayes is not good enough.

Story changes if μ is randomized. For example, *lots* of holes if μ has exponential prior, resampled for each X_n. (In which case X_n has Geometric distribution.

Convergence in mean

- Definition: $X_1, X_2, \ldots, X$ on specified $(\Omega, \mathcal{F}, \mathbb{P})$, finite absolute p^{th} moment. Say $X_n \to X$ in L^p (in p-norm) as $n \to \infty$ if
$$\mathbb{E}\left[|X_n - X|^p\right] \quad \to \quad 0\,.$$
- Motivation: this kind of convergence is highly relevant if one is concerned about risk.
- Remarks:
  - Convergence in p-norm implies convergence in probability.
  - Convergence in probability need not imply convergence in p-norm.
  - Obvious metric distance for p-norm convergence ($p \geq 1$):
  $$d_p(X,Y) = \left(\mathbb{E}\left[|X-Y|^p\right]\right)^{\frac{1}{p}}.$$

1. This can be useful: expectation is linear and so can be easier to handle than probability.
2. The space of random variables with finite p^{th} moment is sometimes represented as L^p or $\mathcal{L}^p$. Strictly L^p should be reserved for *equivalence classes* of random variables (using the equivalence relationship “equal almost surely”).
3. Why $p \geq 1$? Because
(a) then $\|X\|_p = (\mathbb{E}\left[|X|^p\right])^{1/p}$ satisfies a triangle inequality and so defines a *norm* for L^p viewed as the vector space (Banach space) of all random variables of finite p^{th}-moment (treating random variables as equivalent if they agree almost surely); (b) in particular there is a useful duality theory between L^p and L^q when $\frac{1}{p} + \frac{1}{q} = 1$ (if $p = 1$ then take $q = \infty$ and define L^∞ as the family of random variables whose *essential absolute suprema* are finite).
Key words: Minkowski and Hölder inequalities.
4. The case $0 < p < 1$ still makes sense, but we no longer have duality, nor do we have triangle inequalities.
5. Exercise 1.3: (1) convergence in p-norm implies convergence in probability;
(2) counterexamples;
(3) relate p_1-norm and p_2-norm.

Example from Financial Mathematics

- Model financial loss of trader at time n by $Z_n = \exp(X_1 + \ldots + X_n)$, X_i independent $N(-\mu, 1)$, $\mu > 0$.
- It can be shown

$$\frac{X_1 + \ldots + X_n}{n} \quad \to \quad -\mu \qquad \text{almost surely.}$$

Hence $Z_n \to 0$ a.s. (hence also in probability).

- Measure risk of trading by $\mathbb{E}\left[|Z_n - 0|\right] = \mathbb{E}\left[Z_n\right]$.

$$\mathbb{E}\left[Z_n\right] \quad = \quad \left(\mathbb{E}\left[\exp(X_1)\right]\right)^n ,$$

and

$$\mathbb{E}\left[\exp(X_1)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(x - \frac{1}{2}(x+\mu)^2\right) \mathrm{d}\,x = e^{-\mu+\frac{1}{2}} .$$

- If $\mu \leq \frac{1}{2}$ then $Z_1, Z_2, \ldots$ does not converge to 0 in 1-norm.

1. So we are using a conventional and over-simplified log-normal / log-random walk model for the evolution of the loss in time.
2. Argue directly or use SLLN.
3. Averaged quantity of potential loss is taken into account.
4. Indeed if $\mu < \frac{1}{2}$ then $Z_1, Z_2, \ldots$ diverges in 1-norm, even though $Z_n \to 0$ a.s. If $0 < \mu < \frac{1}{2}$ then the loss converges almost surely to zero, but the risk diverges to infinity.
5. The difference between almost sure and 1-norm convergence is very much to the point in the world of finance!
6. **Q:** Under what conditions can 1-norm convergence be implied by convergence in probability?
   A: When *uniform integrability* applies:

$$\lim_{K\to\infty} \sup_n \mathbb{E}\left[|X_n| ; |X_n| > K\right] \quad = \quad 0 .$$

Here $\mathbb{E}\left[|X| ; |X| > K\right]$ is a convenient short-hand notation for the expectation of $|X| \times \mathbb{I}\left[|X| > K\right]$, which is equal to $|X|$ when $|X| > K$ and is zero otherwise. Here is how to see where this definition is coming from: $\mathbb{E}\left[|X|\right] < \infty$ is *equivalent* to the statement $\lim_{K\to\infty} \mathbb{E}\left[|X| ; |X| > K\right] = 0$.

Weak convergence (I)

- Definition: $X_1, X_2, \ldots, X$ not necessarily on same probability space. Say $X_n \to X$ *weakly* (as $n \to \infty$) if, for all bounded continuous functions f,

$$\mathbb{E}\left[f(X_n)\right] \quad \to \quad \mathbb{E}\left[f(X)\right] .$$

- Motivation: Classic example: Binomial Central Limit Theorem. Suppose Y_n is Binomial distribution of parameters n and p, fixed $p \in (0,1)$. Then $X_n = (Y_n - np)/\sqrt{np(1-p)}$ converges weakly to standard Normal distribution.
- Remarks: Equivalent to “convergence of distribution functions” for real-valued X_n.
  Partial converse: “inverse cumulative distribution function” method produces Y_n, Y_∞ defined on same probability space, Y_n, X_n of same distribution, and $Y_n \to Y_\infty$ almost surely.

1. We sometimes write $X_n \Rightarrow X$.
   Functional analysts would prefer to describe this as a kind of weak* convergence.
2. It is easy to forget just how much computation is saved by the simple approximation suggested by this convergence result.
3. Definition above works well for vector-valued random variables, or even more general situations. Polish spaces (complete metrizable spaces) provide a natural context.
4. Partial converse is easiest to prove if distribution function is continuous and is strictly increasing (so inverse is well-defined and continuous). But it is entirely feasible to prove this in general.
5. Exercise 1.5: (1) counterexample;
   (2) inverse c.d.f. method;
   (3) application to representation of weak convergence.

Weak convergence (II)

- There are metrics corresponding to weak convergence. Example: "truncated Wasserstein metric",[2] of importance in optimal transportation theory:

  $$\widetilde{W}(X,Y) = \inf\{\mathbb{E}[1 \wedge |U-V|]\ ;\ U,V \text{ distributed as } X,Y\}\,.$$

- Drop "continuous" from the requirement for f to be "bounded continuous"?

  get *convergence in total variation*: $X_n \to X$ *in total variation* (as $n \to \infty$) if, for all measurable subsets A of the "state-space" (common range of the X_n),

  $$|\,\mathbb{P}[X_n \in A] - \mathbb{P}[X \in A]\,| \quad \to \quad 0\,.$$

  Show we don't need $|\cdot|$. Show this agrees with $\frac{1}{2}\sum_n |p_n - q_n| \to 0$ for discrete random variables. Give example of $X_n \Rightarrow X$ but not $X_n \to X$ in total variation.

[2] "Wasserstein" can also be transliterated "Vasershtein".

1. "Truncated Wasserstein metric" is of importance in optimal transportation theory
2. Convergence in total variation used in Markov chain theory. Mention total variation metric. See also Stein-Chen approximation.
3. Exercise 1.5: (4) total variation;
   (5) total variation for discrete case;
   (6) relate to Wasserstein;
   (7) counterexample.

Weak convergence (III)

- Example: Consider simple random walk $X^{(n)}$ jumping at times $k \times 4^{-n}$, making jumps $\pm 2^{-n}$.

  View piece-wise linear interpolation as random element of *path space* $C([0,1])$.

  Theorem: $X^{(n)}$ converges weakly to Brownian motion.

- Highly useful variant: consider a queue, arrivals according to Poisson process rate λ, i.i.d. service times $\mu T_1, \mu T_2, \ldots$ mean $\mu < \lambda$.

  As $\mu \uparrow \lambda$, so (a re-scaled version of) "work currently in queue" converges weakly to Brownian motion with negative drift reflected at the origin ...

  ... SEE LATER.

We use "Brownian" or "Central Limit Theorem" scaling: time scale factor λ^2, state scale factor λ.

1. Use supremum metric for $C([0,1])$.
2. Reasonable conditions on the distribution of the T_n of course!
3. One of the micro-projects asks you to use simulation and these ideas to investigate a queueing problem motivated by current NHS woes.

Weak Law of Large Numbers

- Theorem (*Weak Law of Large Numbers*, WLLN): Let $X_1, X_2, \ldots$ be i.i.d. on a specified probability space, finite mean μ, finite variance σ^2. As $n \to \infty$,

$$\frac{1}{n}(X_1 + \ldots + X_n) \quad \to \quad \mu \text{ in probability}.$$

- Improvement: "identically distributed" can be replaced by "positive" together with condition controlling size of random variables. Set $m_n = \mathbb{E}[X_1 + \ldots + X_n]$. Then (A) and (B) are *equivalent*:
  - (A) as $n \to \infty$, so $\frac{1}{m_n}(X_1 + \ldots + X_n) \to 1$ in probability, and also $\max_{i=1,\ldots,n} \operatorname{median}(X_i) \to 0$;
  - (B) as $n \to \infty$, so $\sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{m_n} X_i \,;\, \frac{1}{m_n} X_i > \varepsilon\right] \to 0$.

The "jewels" of probability theory are a collection of convergence theorems. I'll discuss WLLN, SLLN, CLT but not LiL.

1. Easily proved using Chebychev inequality and First Borel-Cantelli Lemma. "Independence" can be replaced by "uncorrelated".
2. See Chow and Teicher (2003, §10 Corollary 2)
3. Condition (B) bears a family resemblance to the *Lindeberg condition* arising later for the Central Limit Theorem.
   The improvement provided by Condition (B) is worth bearing in mind. I found it very useful when investigating random scattering processes. And we'll meet a similar notion when discussing CLT.

Strong Law of Large Numbers

- Theorem (*Strong Law of Large Numbers*, SLLN): $X_1, X_2, \ldots$ are *pairwise independent* identically distributed random variables of finite mean μ. As $n \to \infty$,

$$\frac{1}{n}(X_1 + \ldots + X_n) \quad \to \quad \mu \quad \text{almost surely}.$$

- Alternative: *Birkhoff-Khintchine ergodic theorem*: if $X_1, X_2, \ldots$ are *stationary* and *ergodic*, finite mean $\mu = \mathbb{E}[X_n]$, then $\frac{1}{n}(X_1 + \ldots + X_n) \to \mu$ almost surely.
- Transience result: if $X_1, X_2, \ldots$ are jumps of *dependent* random walk with stationary ergodic increments, then $\mu \neq 0$ implies that almost surely the random walk will eventually leave the neighbourhood of 0 for ever. Remarkably, if $\mu = 0$ then dependent random walk will return to any neighbourhood of 0 infinitely often.

1. Proof uses truncation, *via* a Borel-Cantelli argument, Chebyshev's inequality, and some careful analysis.
2. "Pairwise independence" and generalizations ("$(n-1)$-wise independence") have an attractive application to secret-sharing protocols.
3. The recurrence result is essentially an unpublished result of Kesten, Spitzer and Whitman, found in graduate textbooks on probability.
4. Our focus on the independent and identically distributed case should not lead you to suppose that this is the only case for which useful theory exists! Worth considering *why* one might favour the i.i.d. hypothesis . . . .
   The ergodic alternative is worth bearing in mind. I found it very useful when investigating random walk in a random city.

Central Limit Theorem

- Theorem (*Central Limit Theorem*, CLT): $X_1, \ldots, X_n$ i.i.d. finite mean μ, finite variance σ^2. As $n \to \infty$,

$$\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\sigma} \quad \Rightarrow \quad N(0,1)\,.$$

- Improvement: Replace "identical distribution" by *Lindeberg's condition*: if $\mathbb{E}[X_n] = 0$ for all n, and $\operatorname{Var}[X_1] + \ldots + \operatorname{Var}[X_n] = s_n^2 < \infty$, then CLT holds if

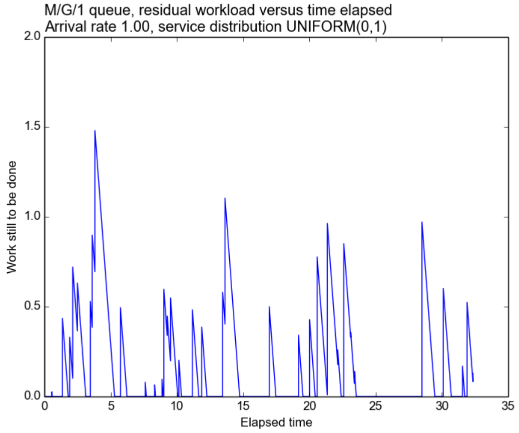
$$\sum_{i=1}^{n} \mathbb{E}\left[\left|\frac{X_i}{s_n}\right|^2 ; \left|\frac{X_i}{s_n}\right| \geq \varepsilon\right] \quad \to \quad 0 \qquad \text{for all } \varepsilon > 0\,.$$

- Alternative: CLT holds even in stationary case, subject to *approximate independence*: (sufficient *mixing* occurs).
- Remark: CLT is "central" because of central rôle. But also "central" because only works in centre of distribution; different methods required for extremes.

1. Nowadays often proved using characteristic functions, but there are other ways.
2. As Le Cam (1986) points out, there are three natural categories for CLT and associates: theorems about normed sums (as above), about triangular arrays, and about approximation theorems. Yes, there are approximation CLTs!
3. Related to WLLN condition $\sum_{i=1}^n \mathbb{E}\left[\frac{1}{m_n}X_i\,;\, \frac{1}{m_n}X_i > \varepsilon\right] \to 0$.
4. Remarkably easy "classical" proof by Lindeberg exchange trick: see Le Cam (1986).
5. Mixing: we need finite **twelfth** moment and α-mixing with $\alpha_n = O(n^{-5})$ (or alternative trade-off between moments and mixing). Here $\alpha_n = \sup\{\mathbb{E}[\phi(X_n)|X_1] : \phi(X_n)$ has mean 0 and unit variance$\}$. Much weaker moment conditions are possible, but mixing conditions must then be strengthened.
Importance of results: demonstrates CLT is very general phenomenon.

Queues in heavy traffic (I)

Residual workload of single-server queue with uniformly distributed service times: scale time, space in same way.



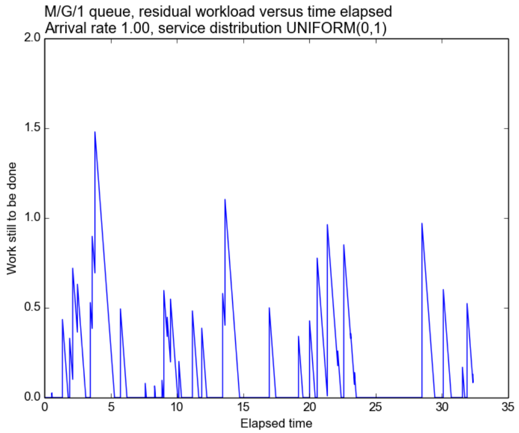
Here we update residual workload W_n at n^{th} arrival as follows:

$$W_n \quad = \quad \min\{W_{n-1} - \text{time since last arrival} + \text{additional work}, 0\}\,.$$

(This presume service occurs at unit rate.)
We see a degenerate limit.

Queues in heavy traffic (II)

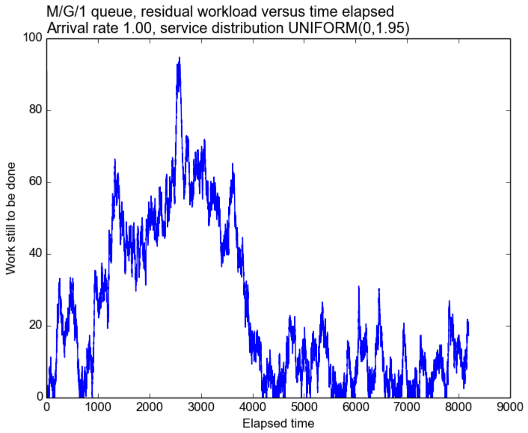
As before, but “CLT scaling”.



“CLT scaling” (*Brownian scaling*) means, scale time by λ^2 and space (here residual workload) by λ.
Note that we still see a degenerate limit!

Queues in heavy traffic (III)

Now use “CLT scaling” and adjust *utilization factor* ρ = arrival rate × mean service time to ensure limiting “constant negative drift” (Asmussen 2003, §VIII.6).

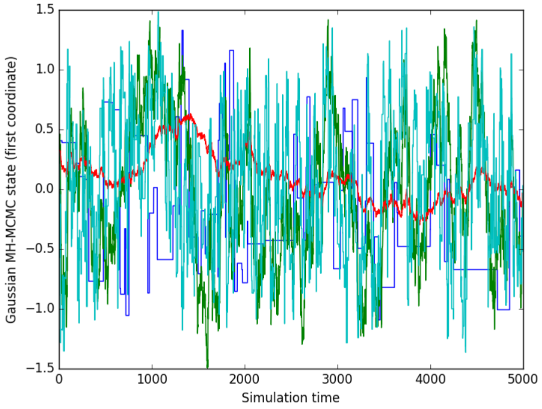


Clear limiting behaviour: in fact the limit is Brownian motion with negative drift, reflected in the origin!
Central limit behaviour: unaffected (subject to regularity) by detailed distribution of service time.
Law of large numbers disposes of variation in inter-arrival times, so (for example) this also works for $GI/G/1$ queues.

Optimal scaling in MCMC

(smooth target, marginal $\propto \exp(-x^4)$)

Target is given by 10 *i.i.d.* coordinates.



Scale parameter for proposal: $\tau = 0.4$ is about right.

Acceptance ratio 26.7%

Example: RW MH-MCMC with Gaussian proposals, target has marginal which is the smooth density

$$\frac{1}{2\Gamma(\frac{5}{4})} \exp(-x^4)\,.$$

This is a high-dimensional ($d = 10$) toy example with independent marginals.

We use a run-length of 5000 steps.

Demonstrates importance of *tuning* (choosing the right scale for the proposal distribution). In fact $\tau = 0.4$ is about right (theory makes sense of what "right" should mean, and uses weak convergence to a specific diffusion!).
Now classical work (Roberts, Gelman, and Gilks 1997): recent work on this by Zanella, Bédard, and WSK (2017).

1. Scale parameter for proposal: $\tau = 1$ is too large! Acceptance ratio 1.7%.
2. Scale parameter for proposal: $\tau = 0.1$ is better. Acceptance ratio 76.5%.
3. Scale parameter for proposal: $\tau = 0.01$ is too small. Acceptance ratio 98.5%.
4. Scale parameter for proposal: $\tau = 0.4$ is pretty much optimal (theory!). Acceptance ratio 26.7%.

Bibliography I

Asmussen, S. (2003).
Applied probability and queues (Second ed.).
New York; Berlin; Heidelberg: Springer.

Chow, Y. S. and H. Teicher (2003).
Probability theory: independence, interchangeability, martingales.
New York - Heidelberg - Berlin: Springer-Verlag.

Grimmett, G. R. and D. Stirzaker (2001).
Probability and random processes.
Oxford University Press.

Bibliography II

Le Cam, L. (1986).
The Central Limit Theorem Around 1935.
Statistical Science 1(1), 78–91.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997).
Weak Convergence and Optimal Scaling of Random Walk Algorithms.
The Annals of Applied Probability 7(1), 110–120.

WSK (1993).
On the empty cells of Poisson histograms.
Journal of Applied Probability 30, 561–574.

Bibliography III

Zanella, G., M. Bédard, and WSK (2017, December).
A Dirichlet form approach to MCMC optimal scaling.
Stochastic Processes and their Applications 127(12), 4053–4082.