Two Elementary Paradoxes

David Aldous

9 March 2019

I maintain a list of 20 open research problems on my web site. These are mostly not in currently active topics. In the abstract I promised to talk about one which is in Tom's general area.

The bond percolation and the SI epidemic models are closely related. In bond percolation on arbitrary finite edge-weighted graphs, the time of emergence of the incipient giant component is (under very weak conditions) deterministic to first order. This corresponds to a pandemic/non-pandemic phase transition in the SI model. Does this very general kind of result extend to the SIS model (contact process) setting?

But this is rather technical and hard to follow in real time – you can read the open problem on the slides to be posted. Instead I will talk about two mathematically elementary "paradoxes," the second being oddly appropriate for this place and time.

Unlike most paradoxes, the first is actually somewhat relevant to a multi-million dollar project. and the second is inspired by (not actually relevant to) an actual multi-billion dollar proposed project.,

Topic 1: Should you do the back-of-an-envelope calculation before the multi-million dollar project?

Non-transitive dice are a "paradox" in the sense that one might just assume such things are impossible without thinking about it. I'll talk about another such paradox.

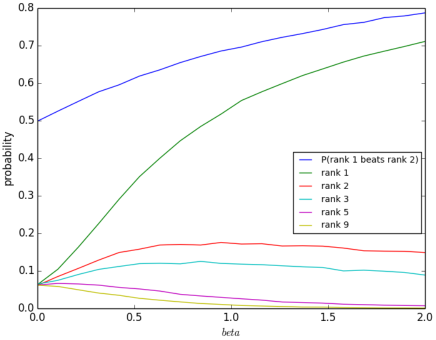
Background analogy: in a sports match the better team doesn't always win, but is likely to win. So in a sports tournament the probabilities of different final winners should be ordered as the abilities of the different teams.

Let's briefly say a math model to check this (details not important). In the Bradley-Terry model

$$\mathbb{P}(A \text{ beats } B) = F(x_A - x_B), \quad x = \text{ ability}, \ F = \text{ logistic}.$$

Make a probability model of random abilities, with a parameter controlling variability of abilities, and simulate a 16 team single-elimination tournament.

Figure: Probabilities of different-ranked players winning the tournament, compared with probability that rank-1 player beats rank-2 player (top curve).



Here math is consistent with common sense.

In a **prediction tournament** contestants state probabilities of future geopolitical events. Here are 5 out of 80 questions asked currently on `gjopen.com`.

• Will an armed group from South Sudan engage in a campaign that systematically kills 1,000 or more civilians during 2019?

• Will there be a new prime minister of the United Kingdom before 1 July 2019?

• Before 1 August 2019, will Facebook announce that Mark Zuckerberg will cease to be the company's sole Chairman or CEO?

• Before 1 October 2019, will the U.S. House of Representatives pass an article of impeachment against President Trump?

• Will China's National People's Congress or its Standing Committee pass a property tax law before 1 October 2019?

• Will North Korea launch an intercontinental ballistic missile (ICBM) before 1 January 2020?

DARPA has a shyer cousin IARPA – non-classified research of indirect interest to the Intelligence community. They funded a series of **Good Judgment Projects** in which volunteers (including me) as individuals and teams make forecasts for such questions.

The point is to gather evidence and expert opinions before giving an answer – and (unlike an exam) there are no limitations – you can copy other people's answers, or if you happen to be a personal friend of Vladimir Putin . . . . . .

Important: contestants are not asked to give a Yes/No prediction, but instead are asked to give a numerical probability, and to update as time passes and relevant news/analysis appears.

Call for one 2018 contest

IARPA is looking for approaches from non-traditional sources that would improve the accuracy and timeliness of geopolitical forecasts. IARPA hosts these challenges in order to identify ways that individuals, academia, and others with a passion for forecasting can showcase their skills easily.

Why Should You Participate: This challenge gives you a chance to join a community of leading experts to advance your research, contribute to global security and humanitarian activities, and compete for cash prizes. This is your chance to test your forecasting skills and prove yourself against the state-of-the-art, and to demonstrate your superiority over political pundits. By participating, you may:

- Network with collaborators and experts to advance your research
- Gain recognition for your work and your methods
- Test your method against state-of-the-art methods
- Win prizes from a total prize purse of $200,000

Why are millions of taxpayer dollars being spent running such projects?

- What makes some individuals better than others? The study starts with a lengthy test of "cognitive style" to see what correlates.
- What makes some teams better than others? How to combine different sources of uncertain information/analysis is a major issue Intelligence assessment. The project managers see team discussions.

How can we assess someone's ability? We do what Carl Friedrich Gauss said 200 years ago – use **mean square error** MSE. An event is a 0 - 1 variable; if we predict 70% probability then our "squared error" is
(if event happens) $(1.0 - 0.70)^2 = 0.09$
(if event doesn't happen) $(0.0 - 0.70)^2 = 0.49$

As in golf, you are trying to get a low score. A prediction tournament is like a golf tournament where no-one knows "par". That is, you can assess people's relative abilities, but you cannot assess absolute abilities.

Writing S for your "tournament score" when the true probabilities of the n events are $(p_i, 1 \leq i \leq n)$ and you predict $(q_i, 1 \leq i \leq n)$,

$$\mathbb{E}S = \sum_i p_i(1 - p_i) + n\sigma^2 \tag{1}$$

where

$$\sigma^2 := n^{-1} \sum_i (q_i - p_i)^2$$

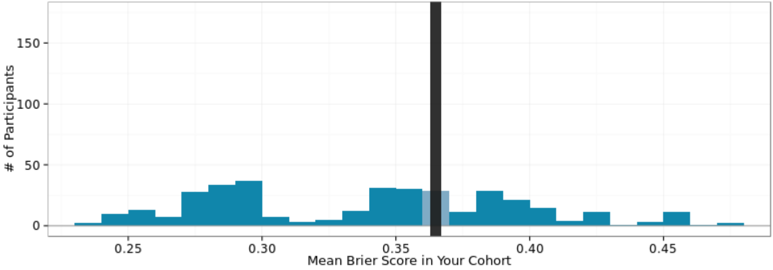
is your MSE (mean squared error) in assessing the probabilities.
So for contestants A and B

$$n^{-1}\mathbb{E}(S_A - S_B) = \sigma_A^2 - \sigma_B^2$$

and so in the long run we can tell who is the more accurate forecaster.

This has philosophical interest, best discussed over beer.

Here is a histogram of 2×scores of individuals in the 2013-14 season `GJP challenge`. The season scores were based on 144 questions, and a back-of-an-envelope calculation gives the MSE due to intrinsic randomness of outcomes as around 0.02, which is much smaller than the spread observed in the histogram. The key conclusion is that there is wide variability between players – as in golf, some people are just much better than others at forecasting these geopolitical events.



In the long run we could tell who is the more accurate forecaster, but what about chance variation in realistic-size tournaments? We need a **model** for comparing contestants scores.

- 100 questions
- true probabilities (unknown to contestants) uniformly spread from 5% to 95%.
- For each contestant A there is a RMS error σ_A for their predicted probabilities: that is, in the model, for each event the prediction $p_{predicted}$ by A is random and such that

$$\sigma_A^2 = \mathbb{E}(p_{predicted} - p_{true})^2.$$

- (complete model specification discussed later)

Now we can simulate the tournament.

Figure: One-on-one comparison: Chance of more accurate forecaster beating less accurate forecaster in 100-question tournament.

		RMS error (less accurate)					
		0.05	0.1	0.15	0.2	0.25	0.3
	0	0.73	0.87	0.95	0.99	1.00	1.00
RMS	0.05		0.77	0.92	0.97	0.99	1.00
error	0.1			0.78	0.92	0.97	0.99
(more	0.15				0.76	0.92	0.97
(accurate)	0.2					0.76	0.91
	0.25						0.73

So this is quite similar to Bradley-Terry: use the RMS probability-prediction error as "ability", and roughly

$$\mathbb{P}(A \text{ beats } B \text{ in prediction tournament }) \approx F(\sigma_A - \sigma_B)$$

for some function F

A leader in this field is Philip Tetlock, with a popular book *Superforecasting* and a 2017 *Science* article and a 2015 paper *Identifying and cultivating superforecasters as a method of improving probabilistic predictions*. They write

> *[the winning strategy for teams over several successive tournaments was] culling off top performers each year and assigning them into elite teams of superforecasters. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics.*

Of course this is essentially the same way that professional football players – or mathematics professors – are developed.

But let's check this holds up mathematically in our prediction context.

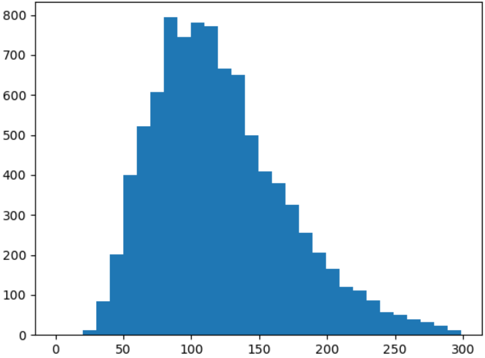
Recall "ability" of contestant formalized as RMS error σ in predicting probability. For a **tournament model** we need a model for variability of ability over contestants:

- 300 contestants
- σ varies evenly from 0 to 0.3.

So we can rank contestants from 1 to 300 in terms of ability. For a tournament with a million events, by LLN the order of scores would closely match the ranking of ability. But what about a realistic size tournament with 100 events?

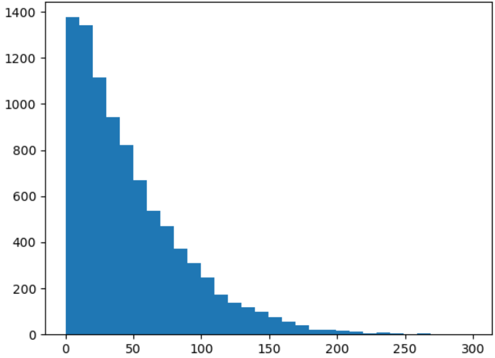
Specifically, what is the (ability) rank of the tournament winner?

Here is the first simulation I did.



Maybe something wrong with my amateur Python code?

Maybe no-one is near-perfect in predicting probabilities? Here are results if the abilities (RMS errors) σ range over $[0.1, 0.4]$ instead of $[0, 0.3]$



This is partly in line with common sense – the best forecasters are relatively more likely to win – but still the winner is liable to be around the 50th best contestant.

So that's the paradox – according to this model, tournaments are a surprisingly ineffective way of identifying the best forecasters, even though IARPA is spending millions of dollars doing precisely this.

Now the issues are

- Is there a calculation to qualitatively explain these simulation model results?
- Why are our model results very different from what is claimed for real tournament results?

The back-of-envelope calculation

Consider a 100-question tournament in which the true probabilities are all 0.5. What are the scores S?

- A perfectly accurate forecaster: $S = 25.0$.
- A contestant who predicts 0.4 or 0.6 randomly on each question: $\mathbb{E}S = 26.0, \ \ \text{s.d.}(S) = 0.98$.
- A contestant who predicts 0.3 or 0.7 randomly on each question: $\mathbb{E}S = 29.0, \ \ \text{s.d.}(S) = 1.83$.

Moreover, as a special feature of the "all true probabilities are 0.5" setting, different contestants' scores are independent. In our simulated setting of 300 contestants, some scores will by chance be around 3 s.d.'s below expectation. With RMS prediction errors ranging from 0 to 0.3, we expect a winning score around 23 and we will not be surprised if this comes from the 100th or 200th best forecaster.

Why is this happening? The key point is that for predicting probabilities the expected cost of small errors scales as $(\text{error})^2$ while the s.d. scales as (error). This is quite different from a typical sport – golf or basketball – where the winner is decided by point difference, points earned in some success/failure way. In sports the expected point difference scales as (difference in ability) and the s.d. of score is roughly constant.

A superficial conclusion of our results is that winning a prediction tournament is strong evidence of superior ability *only* when the better forecasters' predictions are *not* reliably close to the true probabilities. But are our models realistic enough to be meaningful? Two features of our model are unrealistic. One is that contestants have no systematic bias towards too-high or too-low forecasts. But alternate models allowing that give roughly similar results.

I guess the most serious issue is that the errors are assumed independent over both questions and contestants. In reality, if all contestants are making judgments on the same evidence, then (to the extent that relevant evidence is incompletely known) there is surely a tendency for most contestants to be biased in the same direction on any given question. Implicit in our model is that, in a large tournament, this "independence of errors" assumption means that different contestants will explore somewhat uniformly over the space of possible prediction sequences close to the true probabilities, whereas in reality one imagines the deviations would be highly non-uniform.

Statistical analysis of real tournament data is too complicated (for me).
But here are 2 data points.

Note there's also a much deeper philosophical question.

Topic 2: The Shape of Things to Come?

In my "real world" context I make fun of typical math probability papers as "fantasy" – unconnected to any real data. But now I'll tell you an even more extreme fantasy

which will lead to an elementary-to-state math problem.

[David Levinson's *Transportist* blog]

www.citylab.com/perspective/2018/11/transit-city-department-scootershare-ridehail-bikeshare/576982/?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0) [CityLab]

Science

- The Planet Has Seen Sudden Warming Before. It Wiped Out Almost Everything (http://mail01.tinyletterapp.com/DavidLevinson/transportist-january-2019/13301489-t.co/yrpaqsszni?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0). – NY Times
- Are we literally losing our way by relying on GPS devices? Research shows navigating skills do worsen as we depend so much on map apps. (http://mail01.tinyletterapp.com/DavidLevinson/transportist-january-2019/13301493-t.co/il23xhcppv?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0)washingtonpost.com

Fantasy

- Elon Musk's first Boring Company tunnel opens, but the roller-coaster ride has just begun (http://mail01.tinyletterapp.com/DavidLevinson/transportist-january-2019/13301497-t.co/9yzdjggom1?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0) – WaPo

Professoring

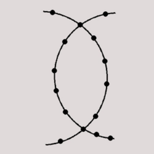
Publishing

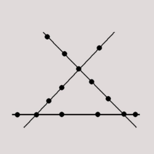
- Heavyweight Showdown Over Research Access: University of California System is playing hardball with Elsevier in negotiations that could transform the way it pays to read and publish research. But does the UC system have the clout to pull it off? (http://mail01.tinyletterapp.com/DavidLevinson/transportist-january-2019/13301501-www.insidehighered.com/news/2018/12/13/university-california-challenges-elsevier-over-access-scholarly-research?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0) – Inside Higher Ed.
- Publish AND perish: how the commodification of scientific publishing is undermining both science and the public good (http://mail01.tinyletterapp.com/DavidLevinson/transportist-january-2019/13301505-transformativelearning.nl/2018/12/04/publish-and-perish-how-the-commodification-of-scientific-publishing-is-undermining-both-science-and-the-public-good/?c=5029e5cd-2f8b-4fc3-9187-6b617b71f2f0) – Transformative Learning

Problem. *Find the connected network of length L that minimizes the expected distance from a random start to the closest point on the network.*

This depends on the density ρ of starting point; as default take 2-dimensional standard Normal.

Background. I have an interest in spatial networks, and the top hit on a Google Scholar search "spatial networks" is a 2011 survey by Marc Barthelemy, a statistical physicist. *Wikipedia – rapid transit* shows typical topologies (shapes) for subway-type networks.





The substantial problem – see Aldous-Barthelemy arXiv paper but not discussed today – is do we reproduce these as optimal under some (slightly) more realistic toy model with several parameters?

Our model today is an "infinite speed and no wait time, no discrete stations" limit model – turns out to be mathematically intriguing in an "your intuition is wrong" sense.

Problem. *Find the connected network of length L that minimizes the expected distance from a random start to the closest point on the network.*

- We seek the actual optimal network for each L – how does the shape evolve as L grows?
- We work numerically. Mostly we consider specific parameterized shapes and optimize over parameters
- Alternatively try simulated annealing to optimize over all networks.

[intuition – board – HSRA]

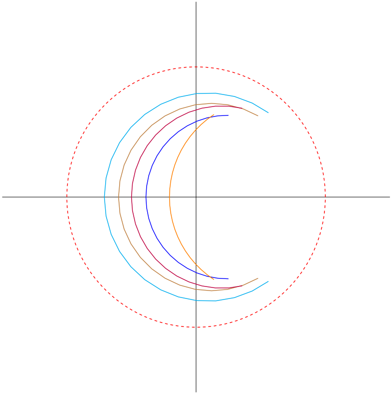
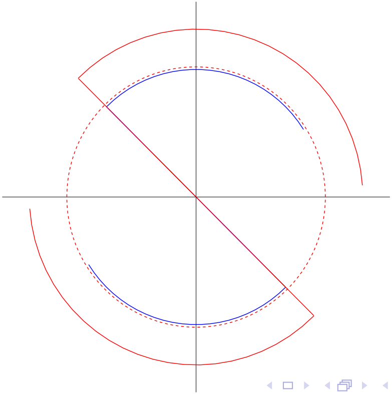


Figure 1: Optimal arcs for L = 1.0, 1.5, 2.0, 2.5, 3.0. The dashed circle is 1 s.d. for the Gaussian population density.

L	5.5	6.5	7.5	8.5	9.5
mean distance	0.544	0.489	0.447	0.418	0.398

Table 2: Data for the S-shape and Gaussian density.



Problem. *Find the connected network of length L that minimizes the expected distance from a random start to the closest point on the network.*

Observation: An optimal network must be a tree (or single path).

Because: If there is a circuit, removing a length ε segment costs order ε^2 but reattaching it elsewhere benefits order ε.

If instead we consider very large networks we do have a theorem concerning the $L \to \infty$ behavior. This result is not so interesting, so let's call it

Take a starting density ρ. Write $d(L)$ for the expected distance-to-network in the optimal network of length L.

Theorem (The Boring Theorem)

$$d(L) \sim \frac{1}{4L} \left(\int_{\mathbb{R}^2} \rho^{1/2}(z)\ dz \right)^2 \text{ as } L \to \infty.$$

What the argument actually shows is that a sequence of networks is asymptotically optimal as $L \to \infty$ if and only if the rescaled local pattern around a typical position z consists of asymptotically parallel lines with spacing proportional to $\rho^{-1/2}(z)$, but the orientations can depend arbitrarily on z. Visualize a fingerprint.

[board – square case]

But surprisingly difficult to see the best finite-L approximation to this limit.

For most values of L the star network with n_L branches of length L/n_L is close to optimal but often not precisely optimal over all shapes.

Indeed by the Boring Theorem it is asymptotically optimal for the non-Gaussian density on the radius-r_0 disc with R uniform on $[0, r_0]$ and θ uniform on $[0, 2\pi]$.

Topic 3: An Open Research Problem.

To me a **network** is a finite (n vertices) connected edge-weighted undirected graph, vertices $v, x, y, \ldots$ and edge weights $w_e = w_{xy}$.

Note two opposite conventions for interpreting weights:

- In TSP-like setting, weight is distance or cost.
- In social networks, weight is strength of relationship (**this talk**).

Tom's 1985 *Interacting Particle Systems* identified several specific processes – Stochastic Ising models, Voter model, contact process, etc – as fundamental. I am more simplistic – to me there are three fundamental random processes one can define over a network:

- the (continuous-time) Markov chain.
- bond percolation
- first passage percolation.

I will discuss bond percolation because it is essentially the SI epidemic model (as the contact process is the SIS epidemic model) and I am interested in what one might be able to say about more realistic epidemic models.

Bond percolation.

An edge e of weight w_e becomes **open** *at an Exponential(w_e) random time.*

In this process we can consider

$$C(t) = \text{max size (number of vertices) in a connected component of open edges at time } t$$

This elates to "emergence of the giant component". Studied extensively on many non-random and specific models of random networks. Can we say anything about $n \to \infty$ asymptotics for (almost) arbitrary networks?

Suppose (after time-scaling) there exist constants $t_* > 0,\ t^* < \infty$ such that

$$\lim_n \mathbb{E} C_n(t_*)/n = 0; \quad \lim_n \mathbb{E} C_n(t^*)/n > 0. \tag{2}$$

In the language of random graphs, this condition says a *giant component* emerges (with non-vanishing probability) at some random time of order 1.

Proposition (1)

Given a sequence of networks satisfying (2), there exist constants τ_n such that, for every sequence $\varepsilon_n \downarrow 0$ sufficiently slowly, the random times

$$T_n := \inf\{t : C_n(t) \geq \varepsilon_n n\}$$

satisfy

$$T_n - \tau_n \to_p 0.$$

The Proposition asserts, informally, that the "incipient" time at which a giant component starts to emerge is deterministic to first order.

Reformulation as epidemics (well known but subtle).

An *SI* model refers to a model in which individuals are either *infected* or *susceptible*. In the network context, individuals are represented as vertices of an edge-weighted graph, and the model is

> *for each edge* (vy), *if at some time one individual* $(v$ *or* $y)$ *becomes infected while the other is susceptible, then the other will later become infected with some* transmission probability p_{vy}.

These transmission events are independent over edges. Regardless of details of the time for such transmissions to occur, this **SI model** is related to the **random graph model** defined by

> *edges* $e = (vy)$ *are present independently with probabilities* $p_e = p_{vy}$.

The relation is:

> *(*) The set of ultimately infected individuals in the SI model is, in the random graph model, the union of the connected components which contain initially infected individuals.*

In modeling an SI epidemic within a population with a given graph structure, we regard edge-weights $w_e = w_{vy}$ as indicating relative frequency of contact. Introduce a *virulence* parameter θ, and define transmission probabilities

$$p_e = 1 - \exp(-w_e\theta). \tag{3}$$

Note this allows completely arbitrary values of (p_e), by appropriate choice of (w_e). Now the point of the parametrization (3) is that the set of potential transmission edges is exactly the same as the time-θ configuration in the bond percolation model. So we can translate our Proposition into a statement about whether the SI epidemic model is pandemic (has $\Theta(n)$ vertices ultimately infected) in terms of the number κ_n of initially infected vertic

Even though this is mathematically trivial, it is **conceptually subtle**. A real-world flu epidemic proceeds in real-world time; instead we just consider the set of ultimately infected people and actual transmission edges; this structure, as a process parametrized by θ, is a nice stochastic process (bond percolation).

Proposition (2)

Take edge-weighted graphs with $n \to \infty$, consider the SI epidemics with transmission probabilities of form (3), and write $C'_{n,\kappa}(\theta)$ for the number of ultimately infected individuals in the epidemic started with κ uniformly random infected individuals. Suppose there exist some $0 < \theta_1 < \theta_2 < \infty$ such that, for all $\kappa_n \to \infty$ sufficiently slowly,

$$\lim_n n^{-1} \mathbb{E} C'_{n,\kappa_n}(\theta_1) = 0; \quad \liminf_n n^{-1} \mathbb{E} C'_{n,\kappa_n}(\theta_2) > 0. \tag{4}$$

Then there exist deterministic $\tau_n \in [\theta_1, \theta_2]$ such that, for all $\kappa_n \to \infty$ sufficiently slowly,

$$n^{-1} C'_{n,\kappa_n}(\tau_n - \delta) \to_p 0, \quad n^{-1} C'_{n,\kappa_n}(\tau_n + \delta) \gg_p 0$$

for all fixed $\delta > 0$.

Proposition 2 provides a subcritical/supercritical dichotomy for the SI epidemics under consideration. The conceptual point is that, for virulence parameter θ not close to the critical value τ_n, either almost all or almost none of the realizations of the epidemic affect a non-negligible proportion of the population. It really is a **phase transition**, and exists for essentially arbitrary large networks.

But the proof is very special. The **open problem** is to prove a similar result for the SIS epidemic (contact process). There are several formulations of plausible conjectures – here is one.

An SIS model (contact process): Given a network (finite connected edge-weighted graph) and a rate function μ_v on vertices v. Introduce a parameter $0 < \theta < \infty$ and a (small) parameter $\varepsilon > 0$.

- Each v is in state S (susceptible) or I (infected); transition rates at v as follows.
- $I \to S$ at rate μ_v.
- $S \to I$ at rate $\varepsilon + \theta \sum \{w_{vy} : \ y \text{ infected } \}$.

Conceptually, you get infected by your contacts with "virulence" parameter θ, or from "outside" with low probability.

Mathematically this is a finite state Markov chain and so has a stationary distribution; we study $X_{\theta,\varepsilon}$ = number of infected vertices, at stationarity.

Now consider a sequence of such networks/rate functions, indexed by $n =$ number of vertices. The basic assumption we will make is: there exist $0 < \theta_* < \theta^* < \infty$ such that, for every sequence $\varepsilon_n \downarrow 0$ sufficiently slowly,

$$n^{-1} X^{(n)}_{\theta_*, \varepsilon_n} \to 0 \text{ in probability}; \quad n^{-1} X^{(n)}_{\theta^*, \varepsilon_n} \gg_p 0. \tag{5}$$

Conjecture

Under assumption (5) (and perhaps further but weak assumptions), there exist $\theta_n \in [\theta_*, \theta^*]$ such that, for all $\varepsilon_n \downarrow 0$ sufficiently slowly,

$$n^{-1} X^{(n)}_{\theta_n - \delta, \varepsilon_n} \to 0 \textit{ in probability}; \quad n^{-1} X^{(n)}_{\theta_n + \delta, \varepsilon_n} \gg_p 0 \ \forall \delta > 0.$$