New Applications of RMT to Statistics and Queueing Theory

Craig A. Tracy

UC Davis

October 2000

I. Queueing Theory & RMT

- P. W. Glynn and W. Whitt
  Annals of Applied Probability, 1991

- Glynn-Whitt/Gravner-Tracy-Widom
  Brownian motion functional
  and universality of *finite* GUE

- K. Johansson's "Shape Fluctuations"
  solution in GUE Universal Regime

- Future Directions?

II. Principal Component Analysis (PCA)

- Crash course in PCA

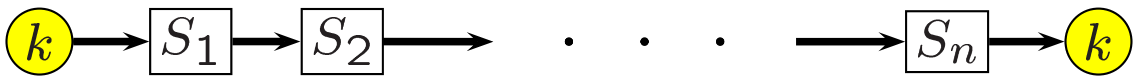
- I. Johnstone
  "On the distribution of the largest
  principal component"

- Future Directions?

Series of n single-server queues each with unlimited waiting space with first in-first out service.

Service times are i.i.d. with mean 1 and variance σ^2.

k customers placed in first queue.



$$\color{blue}{D(k,n)} :=$$

Departure time of customer k from queue n

V(i,j):=Service time for customer i at queue j

Basic Recursion Relation:

$$D(k,n) = \max\{D(k-1,n), D(k,n-1)\} + V(k,n)$$

This implies

$$D(k,n) = \max_{\pi\in\Pi}\left\{\sum_{(i,j)\in\pi} V(i,j)\right\}$$

$\Pi =$ increasing paths $(1,1) \to (k,n)$

Glynn-Whitt Functional Central Limit Theorem

Fix k, number of customers,

$$\frac{D(k,n)-n}{\sigma\sqrt{n}} \xrightarrow{d} \hat{D}_k, \quad n \to \infty$$

$\hat{D}_k$ is a functional of k-dimensional Brownian motion. $\hat{D}_k$ is **independent** of the service-time distribution V. $f = (f_1, f_2, \ldots, f_k) : [0,1] \to \mathbf{R}^k$, continuous, $f(0) = 0$,

$$F(f) \;:=\; \max\{f_1(t_1) + f_2(t_2) - f_2(t_1) + \cdots + f_k(t_k) - f(t_{k-1}) : 0 \le t_1 \le \cdots \le t_k = 1\}$$

$$B = (B_1, \ldots, B_k), k \text{ dim BM}$$

$$\hat{D}_k := F(B)$$

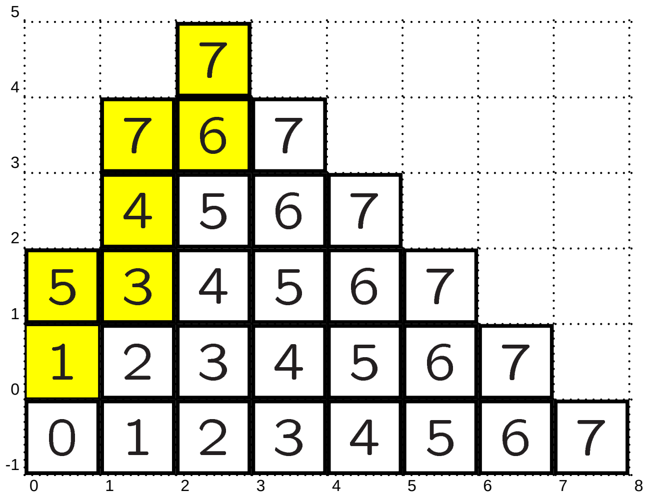
Discrete Time and Space Growth Models
Oriented Digital Boiling
(Gravner-Tracy-Widom)

Introduce a **height function** $h_t(x)$ with rules

$$(i) \quad h_t(x) \leq h_{t+1}(x)$$

$$(ii) \quad \text{If } h_t(x-1) > h_t(x), \text{ then } h_{t+1}(x) = h_t(x-1)$$

$$(iii) \quad \text{Otherwise, then independently of other sites and other times,}$$
$$h_{t+1}(x) = h_t(x) + 1 \text{ with prob } p$$
$$h_{t+1}(x) = h_t(x) \text{ with prob } 1-p.$$



Finite x (=k-1) GUE Regime

Fixed x and $t \to \infty$.

$$\text{Prob}\left(\frac{h_t(x) - p\,t}{(p(1-p)\,t)^{1/2}} < s\right) \xrightarrow{d} F_{x+1,2}(s)$$

$F_{k,2} =$ distr fn for largest eig. in finite k GUE.

GTW also showed

$$\text{Prob}\left(\hat{D}_k \leq s\right) = F_{k,2}(s)$$

Thus the finite GUE distributions are limit laws for a large class of processes; and in this sense are a natural generalization of the gaussian distribution.

Large k and Large n

Glynn-Whitt show (subadditive ergodic theorem) for general service-time distributions having an exponential tail:

$$\frac{D(\lfloor xn \rfloor, n)}{n} \to \gamma(x), \ \text{ a.s. as } n \to \infty$$

where $\gamma(x)$ is a deterministic function depending upon V; for exponentially distributed V

$$\gamma(x) = (1 + \sqrt{x})^2$$

They conclude with asking

$$n^{-\beta}\,(D(\lfloor xn \rfloor, n) - \gamma(x)\,n) \xrightarrow{d} ?$$

This audience realizes that **Kurt Johansson** answered this question for V exponentially distributed. Let

$$\gamma(x) = (1+\sqrt{x})^2, \quad \rho(x) = x^{-1/6}(1+\sqrt{x})^{4/3}$$

Then as $n \to \infty$

$$\text{Prob}\left(\frac{D(\lfloor xn \rfloor, n) - \gamma(x)\, n}{\rho(x) n^{1/3}} \le s\right)$$

$$\begin{aligned} &\to F_2(s) \\ &= \det\left(I - K_{\text{Airy}}\right) \\ &= \exp\left(-\int_s^\infty (x-s) q(x)^2\, dx\right) \end{aligned}$$

$$q = \text{Painlevé II}$$

Future Directions

- We've just seen that the largest eigenvalue distributions of *finite* GUE are limit laws for a large class of processes. Find queueing models in which more general finite RMM arise as limit laws.

- We have a BM functional for the finite x GUE regime. Is there an functional involving infinite dimensional BM for the GUE Universal Regime (F_2)?

- Where do GOE and GSE fit into queueing theory? BM functional?

- What about more general networks of servers? Do RMT techniques have anything to say here?

STATISTICS

Highly recommend

High-dimensional data analysis: The curses and blessings of dimensionality

by **David Donoho**

> The coming century is surely the century of *data*. ...Mathematicians are ideally prepared for appreciating the abstract issues involved in finding patterns in such high-dimensional data.

`http://www-stat.stanford.edu/~donoho/`

Principal Component Analysis

$$\begin{aligned} \mathbf{X} &= \textbf{Data Matrix}, \\ n, \text{Row indices} &= \text{labels of observations} \\ p, \text{Column indices} &= \text{labels of variables} \end{aligned}$$

Climate Studies: n=number of time points and p=number of observation stations

Biology: sample of $n = 40$ winged aphids on which $p = 19$ variables had been measured (body length, body width, . . .).

Search Engines: A common search engine strategy forms huge term (n) by document (p) incident matrices.

The traditional case is

$$\frac{n}{p} \text{ large}$$

but today it is common for both p and n to be large and hence n/p may be moderate to small.

The first part of PCA is simple,

$$S = X^t X = U\,L\,U^t = \sum_{j=1}^{p} \ell_j u_j u_j^t,$$

but why do a PCA?

> Examination of the correlations between the variables in a large data set is often helped by an indication of the weight given to each variable in the first few components. It is then possible to identify groups of variables which tend to vary in much the same way across the set of individuals—reduction of dimensionality by elimination of redundant variables.

Scree plots: One plots the ordered sample eigenvalues and looks for an "elbow" or break between presumably significant and presumably insignificant components, i.e. a focus on the first few largest eigenvalues.

Null hypothesis:

$$X_{i,j} \sim N(0,1), \ \text{ i.i.d.}$$

Define

$$\begin{aligned}
\mu_{np} &= (\sqrt{n-1} + \sqrt{p})^2 \\
\sigma_{np} &= (\sqrt{n-1} + \sqrt{p}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \right)^{1/3}
\end{aligned}$$

Johnstone: If $n/p \to \gamma \geq 1$, then

$$\frac{\ell_1 - \mu_{np}}{\sigma_{np}} \xrightarrow{d} W_1 \sim F_1$$

where

$$F_1(s) = \exp\left(-\frac{1}{2} \int_s^\infty q(x) + (x-s)q(x)^2 \, dx \right)$$

- If $n < p$, reverse roles of n and p.
- Note $p^{1/3}$ behavior of std. dev.

Comments on Proof

- Uses Tracy-Widom GOE/GSE formalism

- That said, lots of estimates need to be proved!

- Some beautiful estimates of Laguerre functions for large order, large variable, and large parameter (α) via Liouville-Green method.

- Complex case ($\beta = 2$) also done. This case previously done by Johansson. This case, though not of as much interest to statisticians, may have applications to signal processing. One gets, of course, F_2 as limiting distribution.

Future Directions

- Data entries X_{ij} are i.i.d. but with "arbitrary" distribution function F. **Soshnikov's** universality theorem for real symmetric Wigner matrices suggests a similar universality theorem here.

- Go beyond i.i.d. case. Statisticians may be more interested in noninvariant measures?

- **Fine tuning** of RMM at the edge produces other universality classes. Will this be of any importance in statistics?

- Johnstone has a number of **empirical observations** where F_1 gives approximate answers. Explore these observations.