Quickest Detection of Changes

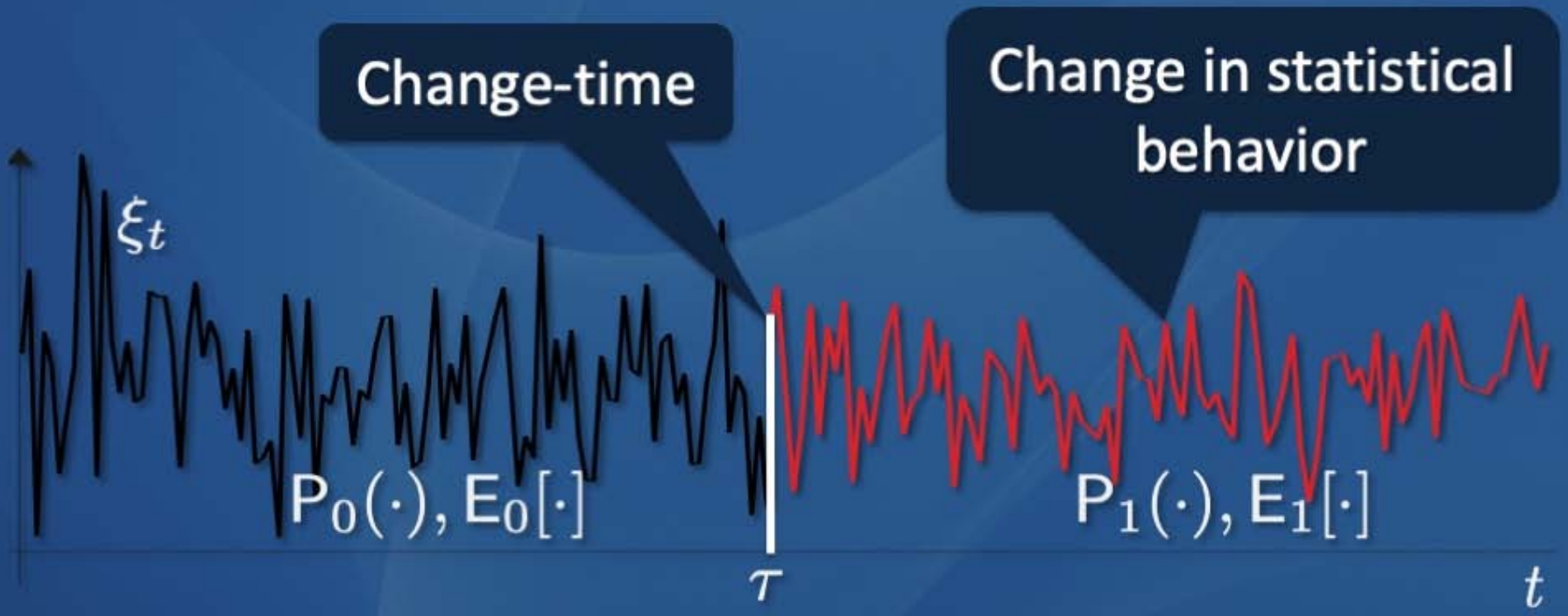
Classical and Modern Formulations

George V. Moustakides
University of Patras, Greece
& Rutgers University, USA

Outline

- Problem definition
- Model for change mechanism
- Performance metrics
- Optimum detectors
- Decentralized formulations
- Data driven version

Problem definition



Detect change as soon as possible

Data become available sequentially: at each time t obtain new sample ξ_t

Detector: Every instant t consult available data $\xi_1, \ldots, \xi_t$ decide if a change took place or not

1 ξ_1 Was there a change? **NO**

2 ξ_1, ξ_2 Was there a change? **NO**

Take more data (continue sampling)

$\vdots$

$\vdots$

T $\xi_1, \xi_2, \ldots, \xi_T$ Was there a change? **YES**

Stop (sampling)

Equivalently

t $\xi_1, \xi_2, \ldots, \xi_t$

Continue or **Stop** sampling

Stopping rule

$G_t(\xi_1, \ldots, \xi_t)$ $\lessgtr$ ν

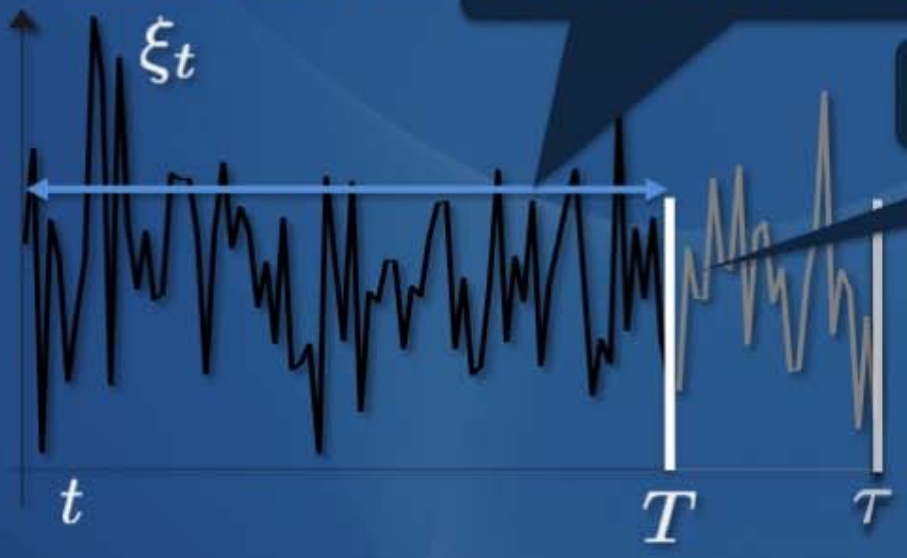
Stop

Continue

Stopping time

$$T = \inf_t \{G_t(\xi_1, \ldots, \xi_t) \geq \nu\}$$

False Alarm



False Alarm Period

Probability of False Alarms

Design T to:
Minimize Detection Delay
or Maximize Detection Probability, while controlling False Alarms

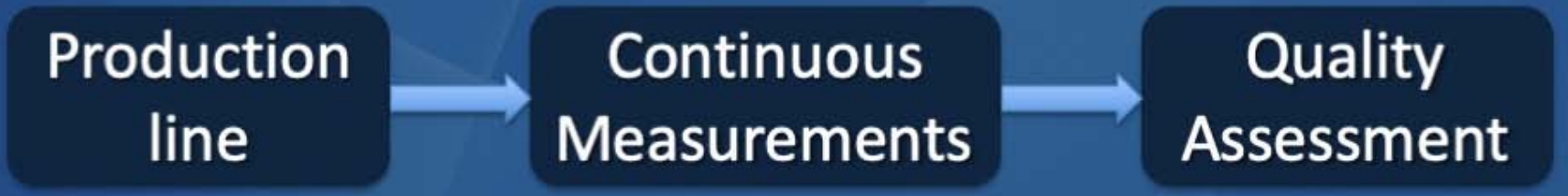
(Successful) Detection



Detection Delay $T - \tau$

Probability of Detection

Quality monitoring of manufacturing process



Medical Applications

Epidemic Detection



Early Detection of Epilepsy Episode



Financial Applications

- Structural Change-detection in Exchange Rates
- Portfolio Monitoring

Electronic Communications
Seismology
Speech & Image Processing (segmentation)
Vibration monitoring (Structural health monitoring)
Security monitoring (fraud detection)
Spectrum monitoring
Scene monitoring
Network monitoring (router failures, attack detection)
⋮

Model for change mechanism



τ

Stopping time

Dependent $\{X_t\}$, $\{\xi_t\}$

X_t: Vibrations at points of the structure

$\xi_t = AX_t + W_t$: Low dimensional noisy observations

If $\|X_t\|^2 \geq \nu$, then crack (change in structure)

Mechanism consults $X_1, \ldots, X_t$ and decides to apply change (**stop** nominal behavior) or not

Quickest detection of changes. ISIT, July 2019, Paris, FRANCE.





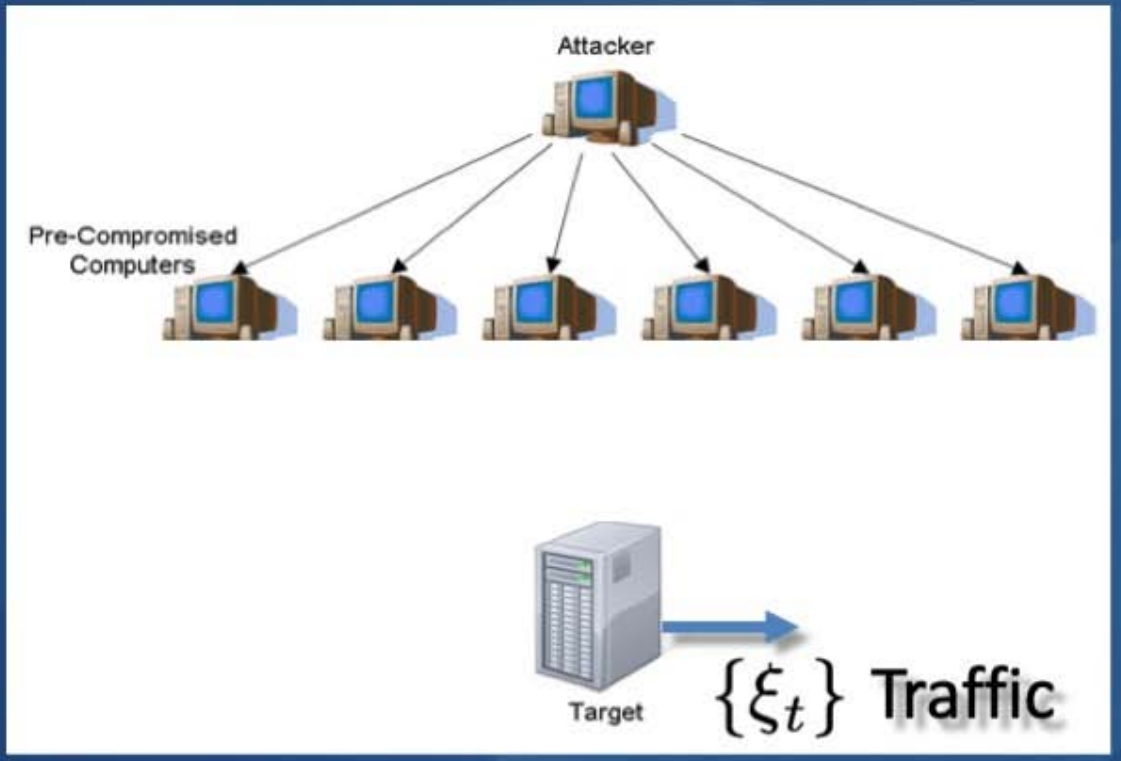


Independent $\{X_t\}$, $\{\xi_t\}$

X_t: Coordinates of ball

ξ_t: Noisy sensor observations

If X_t inside the volume under the goal, "apply" change



At some point in time:
Attack!!!

Attacker no access to observations...

$\{X_t\}, \{\xi_t\}$ independent

Change mechanism
Consults $\{X_t\}$ and controls stopping time τ

Scientist
Consults $\{\xi_t\}$ and controls stopping time T

Design T to detect τ asap

Performance metrics

Delayed Detection

$$\mathcal{J}(T) = \mathsf{E}_1[T - \tau | T > \tau]$$

Hard Limited Delay

$$\mathcal{P}(T) = \mathsf{P}_1(T \leq \tau + M | T > \tau)$$

If stopping rule used by change mechanism unknown computation of $\mathcal{J}(T), \mathcal{P}(T)$ is not possible

We follow a **worst-case analysis**

Delayed Detection

$$\mathcal{J}(T) = \sup_{\tau} \mathsf{E}_1[T - \tau | T > \tau]$$

Lorden (1971)

$$= \sup_{t>0} \sup_{\xi_1,\ldots,\xi_t} \mathsf{E}_1[T - t | T > t, \xi_1, \ldots, \xi_t]$$

$$= \sup_{t>0} \mathsf{E}_1[T - t | T > t] \text{ if } \{X_t\}, \{\xi_t\} \text{ independent}$$

Pollak (1985)

Hard Limited Delay

$$\mathcal{P}(T) = \inf_{\tau} \mathsf{P}_1(T \leq \tau + M | T > \tau)$$

$$= \inf_{t>0} \inf_{\xi_1,\ldots,\xi_t} \mathsf{P}_1(T \leq t + M | T > t, \xi_1, \ldots, \xi_t)$$

$$= \inf_{t>0} \mathsf{P}_1(T \leq t + M | T > t) \text{ if } \{X_t\}, \{\xi_t\} \text{ independent}$$

Delayed Detection

$$\inf_{T} \mathcal{J}(T) =$$

$$\inf_{T} \sup_{t>0} \sup_{\xi_1,\ldots,\xi_t} \mathsf{E}_1[T - t | T > t, \xi_1, \ldots, \xi_t]$$

$$\text{subject to : } \mathsf{E}_0[T] \geq \gamma > 1$$

Hard Limited Delay

$$\sup_{T} \mathcal{P}(T) =$$

$$\sup_{T} \inf_{t>0} \inf_{\xi_1,\ldots,\xi_t} \mathsf{P}_1(T \leq t + M | T > t, \xi_1, \ldots, \xi_t)$$

$$\text{subject to : } \mathsf{E}_0[T] \geq \gamma > 1$$

Optimum detectors

Shiryaev 1963



$\{\xi_t\}$ are i.i.d. before and after the change with corresponding pdfs $\mathrm{f}_0(\xi)$, $\mathrm{f}_1(\xi)$

τ is independent from $\{\xi_t\}$ and follows a geometric distribution $\mathrm{P}(\tau = t) = p(1-p)^t, t = 0, 1, \ldots$

$$S_t = (1 + S_{t-1}) \frac{\mathrm{f}_1(\xi_t)}{(1-p)\mathrm{f}_0(\xi_t)}$$

$$T_{\mathrm{S}} = \inf_{t>0} \{S_t \geq \nu\}$$

Threshold to satisfy false alarm constraint with equality

Lorden 1971



$\{\xi_t\}$ are i.i.d. before and after the change with corresponding pdfs $f_0(\xi)$, $f_1(\xi)$

τ and $\{\xi_t\}$ are dependent

$$\inf_T \mathcal{J}(T) =$$

$$\inf_T \sup_{t>0} \sup_{\xi_1,\ldots,\xi_t} \mathsf{E}_1[T-t|T>t,\xi_1,\ldots,\xi_t]$$

$$\text{subject to : } \mathsf{E}_0[T] \geq \gamma > 1$$

$$S_t = \max\{S_{t-1}, 0\} + \log \frac{f_1(\xi_t)}{f_0(\xi_t)}$$

$$T_C = \inf_{t>0}\{S_t \geq \nu\}$$

Threshold to satisfy false alarm constraint with equality

CUSUM test

CUSUM test

$$S_t = \max\{S_{t-1}, 0\} + \log \frac{f_1(\xi_t)}{f_0(\xi_t)}$$

Known since 1954 as the **Page test**

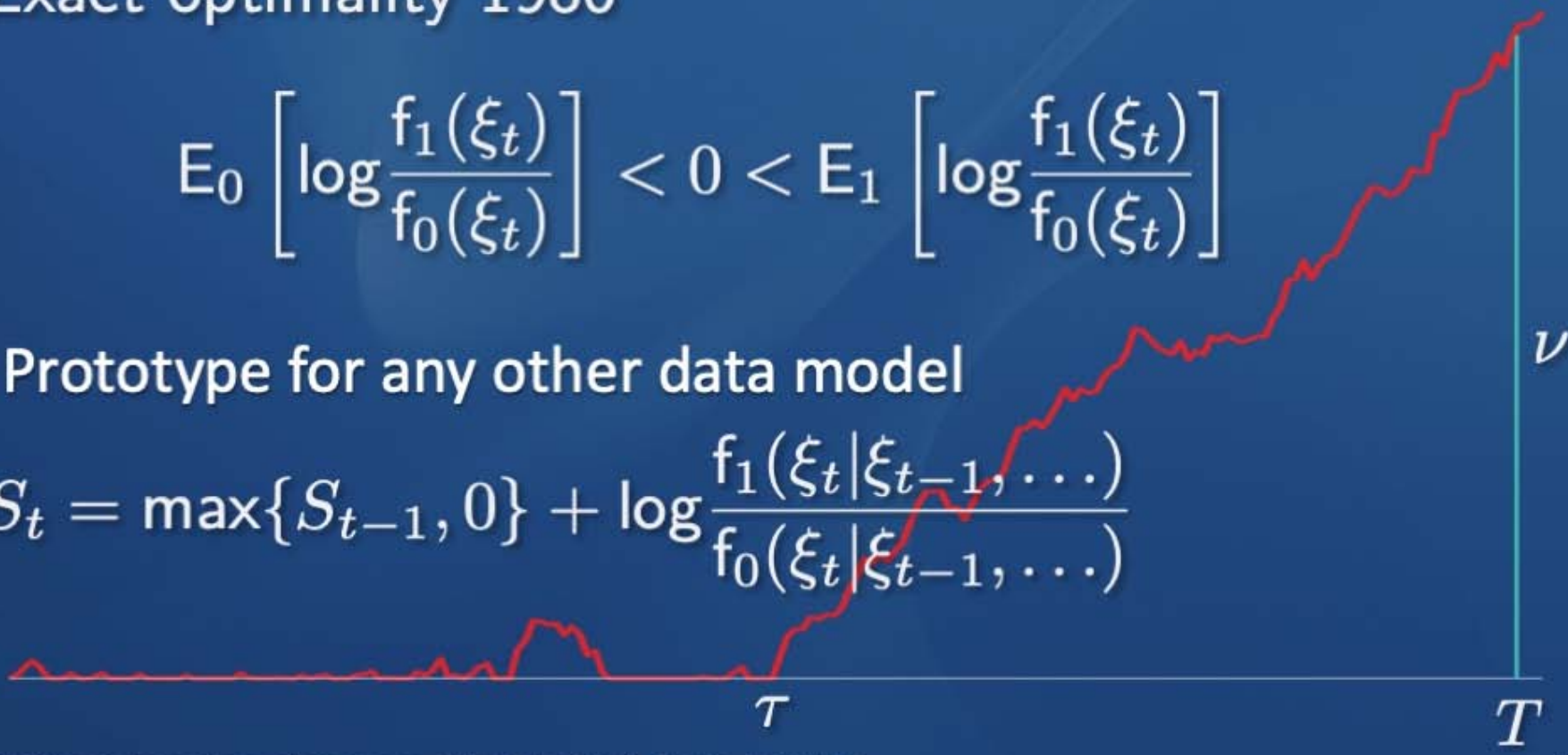
Asymptotic optimality ($\gamma \to \infty$) 1971

Exact optimality 1986

$$E_0\left[\log \frac{f_1(\xi_t)}{f_0(\xi_t)}\right] < 0 < E_1\left[\log \frac{f_1(\xi_t)}{f_0(\xi_t)}\right]$$

Prototype for any other data model

$$S_t = \max\{S_{t-1}, 0\} + \log \frac{f_1(\xi_t|\xi_{t-1}, \dots)}{f_0(\xi_t|\xi_{t-1}, \dots)}$$



2019: $\{\xi_t\}$ are Markov before and after the change with corresponding pdfs $\mathrm{f}_0(\xi_t|\xi_{t-1})$, $\mathrm{f}_1(\xi_t|\xi_{t-1})$

$$S_t = \max\{S_{t-1}, \phi(\xi_{t-1})\} + \log \frac{\mathrm{f}_1(\xi_t|\xi_{t-1})}{\mathrm{f}_0(\xi_t|\xi_{t-1})}$$

$$T_{\mathrm{C}} = \inf_t \{S_t \geq \nu(\xi_t)\}$$

Functions $\phi(\xi), \nu(\xi)$ are solution to a system of integral equations. Computed either numerically or asymptotically $\gamma \to \infty$

As $\gamma \to \infty$ we have $\phi(\xi) \to 0$ and $\nu(\xi) \to \nu$

Pollak 1985

$\{\xi_t\}$ are i.i.d. before and after the change with corresponding pdfs $\mathrm{f}_0(\xi)$, $\mathrm{f}_1(\xi)$

τ and $\{\xi_t\}$ are independent

$$\inf_T \mathcal{J}(T) = \inf_T \sup_{t>0} \mathrm{E}_1[T-t|T>t]$$

$$\text{subject to : } \mathrm{E}_0[T] \geq \gamma > 1$$

Mei 2006

Shiryaev-Roberts-Pollak test

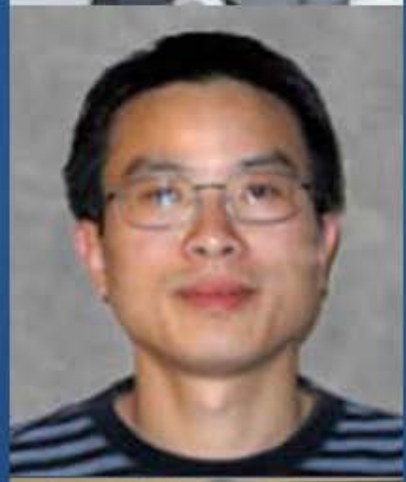
$$S_t = (S_{t-1}+1)\frac{\mathrm{f}_1(\xi_t)}{\mathrm{f}_0(\xi_t)} \quad T_{\mathrm{SRP}} = \inf_t \{S_t \geq \nu\}$$

If S_0 specially designed T_{SRP} asymptotically ($\gamma \to \infty$) optimum.

Tartakovsky 2010

Exact optimality? 1997-2006







Tartakovsky 2019



τ and $\{\xi_t\}$ are independent

$$\inf_T \mathcal{J}(T) = \inf_T \sup_{t>0} \mathsf{E}_1[T-t|T>t]$$

$$\text{subject to : } \mathsf{E}_0[T] \geq \gamma > 1$$

Multiple post-change possibilities. Data after change i.i.d. with pdf $\mathsf{f}_1(\xi), \ldots, \mathsf{f}_k(\xi)$

$$S_t^i = (S_{t-1}^i + 1)\frac{\mathsf{f}_i(\xi_t)}{\mathsf{f}_0(\xi_t)}, \ i = 1, \ldots, k$$

$$T_{\mathrm{T}} = \inf_{t>0}\{S_t^1 + \cdots + S_t^k \geq \nu\}$$

T_{T} asymptotically ($\gamma \to \infty$) optimum

Hard Limited Delay

τ and $\{\xi_t\}$ are dependent

$$\sup_T \mathcal{P}(T) = \sup_T \inf_{t>0} \inf_{\xi_1,\ldots,\xi_t} \mathsf{P}_1(T \le t + M | T > t, \xi_1, \ldots, \xi_t)$$

$$\text{subject to : } \mathsf{E}_0[T] \ge \gamma > 1$$

τ and $\{\xi_t\}$ are independent

$$\sup_T \mathcal{P}(T) = \sup_T \inf_{t>0} \mathsf{P}_1(T \le t + M | T > t)$$

$$\text{subject to : } \mathsf{E}_0[T] \ge \gamma > 1$$

If $\{\xi_t\}$ are i.i.d. before and after the change with pdfs $\mathsf{f}_0(\xi)$, $\mathsf{f}_1(\xi)$, and interested in $M = 1$ (detect immediately)

$$T_{\mathrm{Sh}} = \inf_{t>0} \left\{ \frac{\mathsf{f}_1(\xi_t)}{\mathsf{f}_0(\xi_t)} \ge \nu \right\}$$

Shewhart test (1931)

Markovian pre- and post-change pdfs for observations

$$\sup_T \mathcal{P}(T) = \sup_T \inf_{t>0} \inf_{\xi_1,\ldots,\xi_t} \mathsf{P}_1(T = t+1 | T > t, \xi_1, \ldots, \xi_t)$$

$$\text{subject to : } \mathsf{E}_0[T] \geq \gamma > 1$$

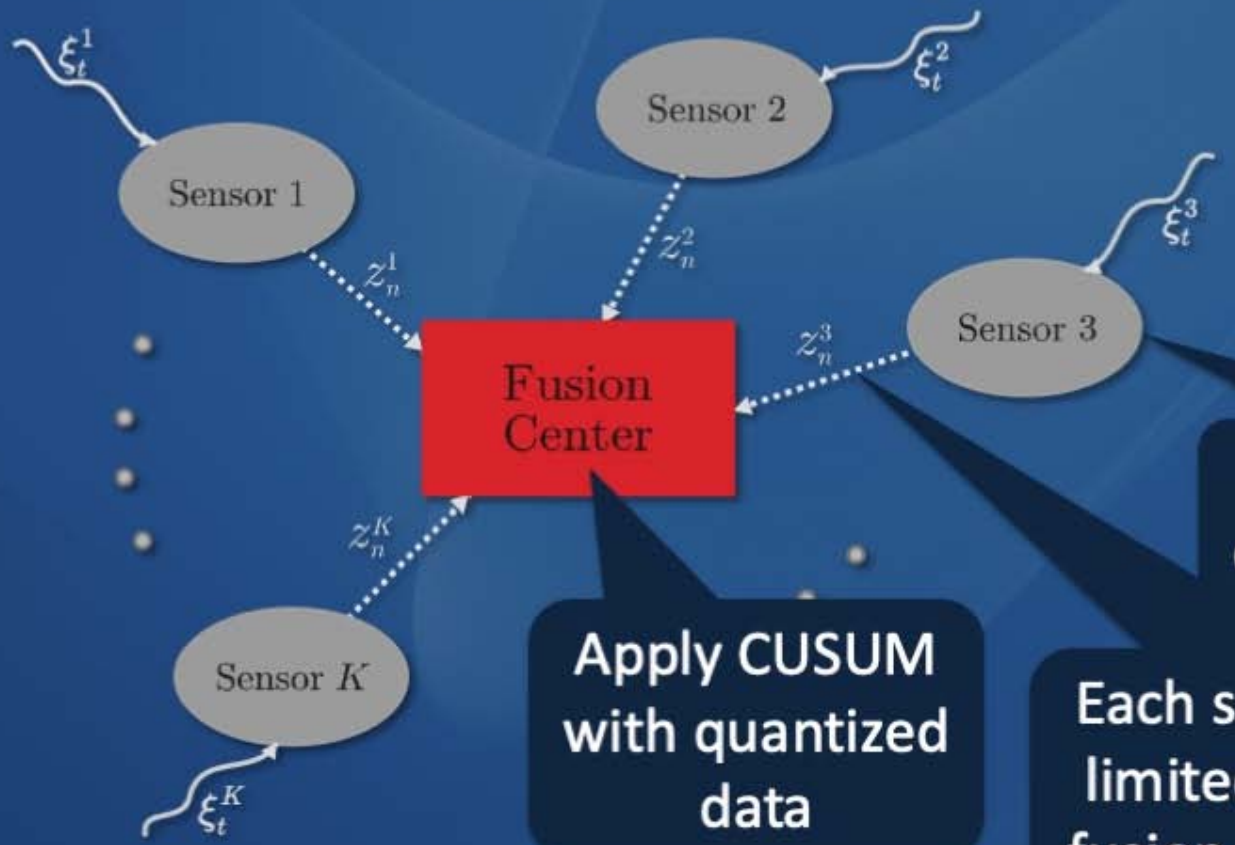
$$T_{\mathrm{Sh}} = \inf_{t>0} \left\{ c(\xi_{t-1}) \frac{\mathsf{f}_1(\xi_t | \xi_{t-1})}{\mathsf{f}_0(\xi_t | \xi_{t-1})} \geq \nu(\xi_t) \right\}$$

2015: Functions $c(\xi), \nu(\xi)$ satisfy system of integral equation. Can be computed numerically or asymptotically ($\gamma \to \infty$). Simple solution for conditionally Gaussian: $\xi_t = \alpha(\xi_{t-1}) + w_t$ where $\{w_t\}$ i.i.d. $\mathcal{N}(0,1)$

Decentralized formulations

Veeravalli 2001





Simultaneous change in all sensors

Apply CUSUM with quantized data

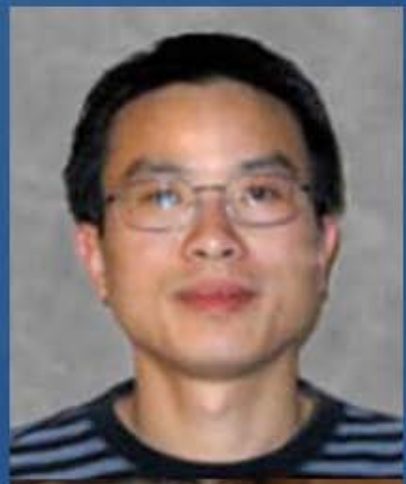
Each sensor must send limited information to fusion center (i.e. 1 bit)

Veeravalli 2001: Apply optimum quantization on each likelihood ratio $\frac{f_1^i(\xi_t)}{f_0^i(\xi_t)}$

Mei 2011 - Fellouris 2016



There is a change in statistical behavior of a subgraph.





$$S_t^i = \max\{S_{t-1}^i, 0\} + \log \frac{f_1^i(\xi_t)}{f_0^i(\xi_t)}$$

$$T = \inf_{t>0}\{\max_i S_t^i \geq \nu\}$$

$$T = \inf_{t>0}\{S_t^1 + \cdots + S_t^K \geq \nu\}$$

$$T = \inf_{t>0}\left\{ \sum_{i=1}^K \log\left(1 - \pi + \pi e^{S_t^i}\right) \geq \nu\right\}$$

Data driven version (ongoing)

What if $f_0(\xi), f_1(\xi)$ **NOT known**

$\{\xi_1^0, \ldots, \xi_N^0\}$ pre-change (training) data

$\{\xi_1^1, \ldots, \xi_N^1\}$ post-change (training) data

Use data to estimate $f_0()$,$f_1()$

Density estimation, not very efficient

$$S_t = \max\{S_{t-1}, 0\} + \log \frac{f_1(\xi_t)}{f_0(\xi_t)}$$

Can we train a NN to estimate the $\log \frac{f_1(\xi_t)}{f_0(\xi_t)}$ using the training data ?

Need proper optimization problem!!!

Consider scalar functions $\phi(z), \psi(z)$ and scalar function $u(\xi)$. Define average cost:

$$\mathcal{J}(u) = \mathsf{E}_0\left[\phi\big(u(\xi)\big)\right] + \mathsf{E}_1\left[\psi\big(u(\xi)\big)\right]$$

If we want $\min_u \mathcal{J}(u)$ to yield

$$u(\xi) = \log \frac{\mathsf{f}_1(\xi)}{\mathsf{f}_0(\xi)}$$

it is necessary and sufficient

$$\psi'(u) = -e^{-u}\phi'(u)$$

$$\mathcal{J}(u) = \mathsf{E}_0\left[\phi\big(u(\xi)\big)\right] + \mathsf{E}_1\left[\psi\big(u(\xi)\big)\right]$$

Let $u(\xi, \theta)$ be the output of a NN with input ξ then

$$\hat{\mathcal{J}}(\theta) = \tfrac{1}{N}\sum_{l=1}^{N} \phi\big(u(\xi_l^0, \theta)\big) + \tfrac{1}{N}\sum_{l=1}^{N} \psi\big(u(\xi_l^1, \theta)\big)$$

Optimize over θ (training)

$$\log \frac{\mathsf{f}_1(\xi_t|\xi_{t-1})}{\mathsf{f}_0(\xi_t|\xi_{t-1})} \text{ ???}$$

$$\theta_t = \theta_{t-1} - \mu\left\{\nabla_\theta \phi\big(u(\xi_t^0, \theta_{t-1})\big) + \nabla_\theta \psi\big(u(\xi_t^1, \theta_{t-1})\big)\right\} \to \theta_*$$

$$S_t = \max\{S_{t-1}, 0\} + u(\xi_t, \theta_*) \qquad T_{\mathrm{C}} = \inf_{t>0}\{S_t \geq \nu\}$$

2019 IEEE International Symposium on Information Theory

7-12 July 2019 • Paris, France

Plenary Speakers

Monday, 8 July, 08:30	Muriel Médard
Tuesday, 9 July, 08:30	Alice Guionnet
Wednesday, 10 July, 08:30	Shannon Lecturer Erdal Arıkan
Thursday, 11 July, 08:30	Yoshua Bengio
Friday, 12 July, 08:30	George V. Moustakides

Shannon Lecturer

Erdal Arıkan — Wednesday, 10 July, 08:30

From Sequential Decoding to Channel Polarization and Back Again



Abstract

This talk gives an account of the original ideas that motivated the development of polar coding and discusses some new ideas for exploiting channel polarization.

Biography

Erdal Arıkan was born in Ankara, Turkey, in 1958. He received the B.S. degree from the California Institute of Technology, Pasadena, CA, in 1981, and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1982 and 1985, respectively, all in Electrical Engineering. He served as an assistant professor at the University of Illinois at Urbana-Champaign during 1986-1987. Since Sept. 1987, he has been with the Electrical-Electronics Engineering Department of Bilkent University, Ankara, Turkey, where he works as a Professor. He is the recipient of the 2010 IEEE Information Theory Society Best Paper Award, the 2013 IEEE W. R. G. Baker Award, IEEE Turkey Section 2017 Life-Long Achievement Award, and the 2018 IEEE Hamming Medal. Arıkan is a member of the IEEE and an IEEE Fellow.

Plenary Speakers

Muriel Médard — Monday, 8 July, 08:30

Le jeu du réseau et du hasard - exploiting randomness

Abstract



In this talk, we consider two different aspects of randomness as a design principle. In the first part, we overview the intentional use of randomness for network code constructions. Random linear network coding (RLNC) allows for endless composability and, in erasure networks, achieves network capacity. In the second part of the talk, we exploit the randomness of noise in channels for decoding. Guessing random additive noise to decode (GRAND) does not use the structure of the code, but instead considers the random properties of the noise itself. GRAND provides a maximum a posteriori decoder with any code, with attractive complexity properties for many channels.

Biography

Muriel Médard is the Cecil H. Green Professor in the Electrical Engineering and Computer Science (EECS) Department at MIT and leads the Network Coding and Reliable Communications Group at the Research Laboratory for Electronics at MIT. She has co-founded three companies to commercialize network coding, CodeOn, Steinwurf and Chocolate Cloud. She has served as editor for many publications of the Institute of Electrical and Electronics Engineers (IEEE), of which she was elected Fellow, and she has served as Editor in Chief of the IEEE Journal on Selected Areas in Communications. She was President of the IEEE Information Theory Society in 2012, and served on its board of governors for eleven years. She has served as technical program committee co-chair of many of the major conferences in information theory, communications and networking. She received the 2009 IEEE Communication Society and Information Theory Society Joint Paper Award, the 2009 William R. Bennett Prize in the Field of Communications Networking, the 2002 IEEE Leon K. Kirchmayer Prize Paper Award, the 2018 ACM SIGCOMM Test of Time Paper Award and several conference paper awards. She was co-winner of the MIT 2004 Harold E. Edgerton Faculty Achievement Award, received the 2013 EECS Graduate Student Association Mentor Award and served as Housemaster for seven years. In 2007 she was named a Gilbreth Lecturer by the U.S. National Academy of Engineering. She received the 2016 IEEE Vehicular Technology James Evans Avant Garde Award, the 2017 Aaron Wyner Distinguished Service Award from the IEEE Information Theory Society and the 2017 IEEE Communications Society Edwin Howard Armstrong Achievement Award.

Alice Guionnet — Tuesday, 9 July, 08:30

Rare events in Random Matrix Theory



Abstract

The typical behaviour of random matrices is well understood in a rather large range of models, including deformed models of a signal+a noise. Estimating the probability of deviating from these typical behaviours is a much more challenging question that we shall discuss in this talk.

Biography

Alice Guionnet entered Ecole Normale superieure in 1989. She earned her PhD in 1995 in probability theory under the supervision of Gerard Ben Arous ,working on spin glass dynamics and aging. She is well known for her work on random matrices. She has established surprizing links with various other fields of mathematics as spectral theory, operator algebra, free probability which lead her to several outstanding results. Her "single ring theorem" is a real masterpiece of analysis. One of the most important contribution of Alice Guionnet might be a series of work where she founds the theory of "Matrix Models". She has received a number of prestigious prizes like the Loeve prize, the silver medal of CNRS and the Blaise Pascal medal, showing her impressive impact beyond probability theory. After being an invited speaker at the International Congress of Mathematician, and at the International Congress of Mathematical Physics, she was elected in 2017 at the french Academy of Sciences.

Yoshua Bengio — Thursday, 11 July, 08:30

Information Theory for and by Deep Learning



Abstract

One of the central goals of deep learning is to learn good representations, which ideally disentangle the underlying abstract explanatory factors of the observed data. How goodness of these representations should be defined is an active and open area of research, but an interesting hypothesis is that we should try to define this

mostly in the high-level representation space itself, rather than in data space, because the interesting factors (e.g., the word sequence, in speech) may represent very few bits out of the actual sensory signal (e.g. the acoustic speech signal). Yet, standard machine learning objectives (like maximum likelihood) are defined in the data space. This has stimulated the exploration of training objectives based on information-theoretical ideas, like mutual information and entropy. Interestingly, the tools of deep learning, in particular of adversarial networks, is being used to learn these objective functions: the basic idea is that dependency can be measured by how well a classifier can separate samples from a joint distribution from samples of the marginals. Whereas classical non-parametric methods struggle to estimate quantities like entropy or mutual information when the variables are high-dimensional, a new wave of estimators has been proposed based on neural networks, which also raise other interesting questions and open new opportunities for learning high-level representations.

Biography

Recognized as one of the world's leading experts in artificial intelligence and a pioneer in deep learning, Yoshua Bengio studied in Montreal, earned his Ph.D. in computer science from McGill University in 1991, and did post-doctoral studies at MIT.

Since 1993, he has been a professor in the Department of Computer Science and Operations Research at the Université de Montréal, and he holds the Canada Research Chair in Statistical Learning Algorithms. In addition, he is Scientific Director of IVADO and Mila, the Quebec Artificial Intelligence Institute, the world's largest deep learning academic research group.

An Officer of the Order of Canada, he is also a Fellow of the Royal Society of Canada, the recipient of the Marie-Victorin Prize in 2017, and was named Radio-Canada's Scientist of the Year for 2017. In 2018, he was awarded the 50th anniversary medal of Quebec's Ministry of International Relations and La Francophonie.

Yoshua Bengio is one of the world's most cited computer scientists, thanks to his three books and more than 500 publications. His h-index stands at 131, with more than 149,000 Google Scholar citations. His ambition is to understand the principles that lead to intelligence through learning, as well as promote the development of artificial intelligence for the benefit of all.

George V. Moustakides

Friday, 12 July, 08:30

Quickest Detection of Changes: Classical and Modern Approaches



Abstract

The problem of rapid detection of a change in the statistical behavior of an observed data sequence, finds application in a plethora of scientific fields. Quality control, detection of seismic wave onset time, epidemic detection, portfolio monitoring, health monitoring of structures, fraud detection, spectrum monitoring, attack and router failure detection in networks, are only a few examples where quickest (sequential) change detection can be adopted to mathematically formulate the corresponding application problem. In the first part of our presentation we introduce the notion of a stopping time which is the mathematical entity we employ for implementing a sequential detector. We continue with an effort to understand the various change imposing mechanisms that exist in nature and we provide a high-level model for their statistical description. This in turn allows us to define proper performance measures and related optimization problems which, when solved, give rise to optimum detection strategies. In the second part, we make an overview of the most popular metrics proposed in the literature along with their optimum detectors and discuss their applicability to real problems. Finally, in the last part of our presentation we focus on modern versions of the quickest change detection problem. In particular, we consider data that are acquired from multiple sources with the change occurring simultaneously either in all or in an unknown subset of the nodes. We also consider different classes of sequential detectors as centralized, decentralized, and the more recently investigated, distributed and discuss their corresponding optimality properties.

Biography

George V. Moustakides received the diploma in Electrical and Mechanical Engineering from the National Technical University of Athens, Greece in 1979, the MSE in Systems Engineering from the University of Pennsylvania in 1980, and the M.Sc and Ph.D in Electrical Engineering and Computer Science from Princeton University in 1983. Since 1988 he is with the University of Patras, Greece initially with the Computer Engineering and Informatics department and afterwards with the department of Electrical and Computer Engineering. In 2017 he also joined the Computer Science department at Rutgers University. In the past he held various visiting or long-term appointments with INRIA, Princeton University, University of Pennsylvania, Columbia University, University of Maryland, Georgia Institute of Technology, University of Southern California, University of Illinois at Urbana-Champaign and Rutgers University. His interests include Sequential Detection, Statistical Signal Processing and Machine Learning. From 2011 to 2014 and from 2016 to 2018 he served as Associate Editor for the IEEE Transactions on Information Theory.