Why big data and machine learning must meet fractional calculus?

YangQuan Chen, Ph.D., Director, MESA LAB (Mechatronics, Embedded Systems and Automation) ME/EECS/SNRI/HSRI/CITRIS, School of Engineering, University of California, Merced

E: yqchen@ieee.org; or, yangquan.chen@ucmerced.edu
T: (209)228-4672; O: SRE-327; Lab: Castle #22 (T: 228-4398)

July 6th, 2020. Monday 2-3PM
Seminar over internet hosted by Beijing Jiaotong University
Acknowledgements

• Professor Yongguang Yu and BJTU

• You all, for coming!

• Am I nervous? - *Fractionally*
Skip Ad in a *fractional* hour
University of California, Merced

- The Research University of the Central Valley
- Centrally Located
  - Sacramento – 2 hrs
  - San Fran. – 2 hrs
  - Yosemite – 1.5 hrs
  - LA – 4 hrs
- Surrounded by farmlands and sparsely populated areas

07/06/2020 Why BD and ML must meet FC?
UC Merced

- Established 2005
- 1st research university in 21st century in USA.
- 6815/7,375 Undergraduates
- 521/592 Grads (most Ph.Ds)
- 60% (70%) 1st generation; 60% Pell
- Strong Undergraduate Research Presence (HSI, MSI)
- 2020: 9K undergrads, 1K grads
- $1.3B expansion: now ~ 2020

http://www.ucmerced.edu/fast-facts

07/06/2020 Why BD and ML must meet FC?

<table>
<thead>
<tr>
<th>Campus</th>
<th>Public</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC Berkeley</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>UCLA</td>
<td>2 / 1</td>
<td>24</td>
</tr>
<tr>
<td>UC Santa Barbara</td>
<td>8</td>
<td>37</td>
</tr>
<tr>
<td>UC Irvine</td>
<td>9</td>
<td>39</td>
</tr>
<tr>
<td>UC Davis</td>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>UC San Diego</td>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>UC Santa Cruz</td>
<td>30</td>
<td>79</td>
</tr>
<tr>
<td>UC Riverside</td>
<td>56</td>
<td>118</td>
</tr>
<tr>
<td>UC Merced</td>
<td>78 / 67 / 44</td>
<td>152 / 136 / 107</td>
</tr>
</tbody>
</table>

[1](https://www.universityofcalifornia.edu/news/6-uc-campuses-named-among-nation-s-top-10-public-universities)

Why BD and ML must meet FC?
### 2020 U.S. News and World Report Rankings
**Graduate / Engineering**

<table>
<thead>
<tr>
<th>Campus</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC Berkeley</td>
<td>3</td>
</tr>
<tr>
<td>UC San Diego</td>
<td>11</td>
</tr>
<tr>
<td>UCLA</td>
<td>16</td>
</tr>
<tr>
<td>UC Santa Barbara</td>
<td>24</td>
</tr>
<tr>
<td>UC Davis</td>
<td>31</td>
</tr>
<tr>
<td>UC Irvine</td>
<td>36</td>
</tr>
<tr>
<td>UC Riverside</td>
<td>75</td>
</tr>
<tr>
<td>UC Santa Cruz</td>
<td>87</td>
</tr>
<tr>
<td>UC Merced</td>
<td>134/107</td>
</tr>
</tbody>
</table>

[Link](https://www.universityofcalifornia.edu/news/thinking-about-graduate-school-uc-programs-top-us-news-2020-rankings) Why BD and ML must meet FC?
Mechatronics, Embedded Systems and Automation Lab

Real solutions for sustainability!

Established Aug. 2012 @ Castle, 4,500+ sq ft
6 Ph.D/10+ undergraduate researchers
10+ visiting scholars || sponsored / mentored many capstone teams

Education and Outreach Activities:
- Eng Service Learning (Sp14)
- AIAA Student Branch @UCM
- Preview Days, Bobcat Day etc.
- Robots-n-Ribs|MESABox! STEM-TRACKS TEAM-E; UAS4STEM. USDA HSI: 2016-20
- ME142 Mechatronics (take-home labs)
- ME280 Fractional Order Mechanics
- ME211 Nonlinear Control
- ME143 Unmanned Aerial Systems
- ME212 Robustness and Optimality

Research Areas of Excellence:
(ISI H-index=61, Google H=83; i10=463, HCR-2018,19)
- Unmanned Aerial Systems & UAV-based Personal Remote Sensing (PRS)
- Cyber-Physical Systems (CPS)
  - Mechatronics
  - Applied Fractional Calculus
  - Modeling and Control of Renewable Energy Systems

Projects Related to San Joaquin Valley:

Energy [Solar/wind energy, Building efficiency (HVAC lighting), smart grids integration, NG pipelines]
- Water (Water/soil salinity management, water sampling UAVs)
- Precision Ag/ Environment (Crop dynamics, optimal harvesting, pest, methane sniffing/mapping, DH ...)
UC Multi-campus Synergy on CIDERS
California Institute of Data-drone Engineering and Services

UCM, UCSC, UCB, UCSD, LLNL

CIDERS in Scientific data-drones: platforms, operation, and certification

CIDERS in precision agriculture

CIDERS in environmental monitoring: water, fire, soil, dust, AQ …

Why BD and ML must meet FC?
Outline

• Fractional Calculus, Complexity, and Fractional Order Thinking
• Big Data, Variability, and Fractional Calculus
• Machine Learning, Optimal Randomness and Fractional Calculus
• Looking Into Future: Fractional Calculus is Physics Informed
What is “Fractional Calculus”?

• **Calculus**: integration and differentiation.

• “**Fractional Calculus**”: integration and differentiation of non-integer orders.
  – Orders can be real numbers (and even complex numbers!)
  – Orders are not constrained to be “integers” or even “fractionals”

**How this is possible?**

**Why should I care?**

**Any (good) consequences (to me)?**

Why BD and ML must meet FC?
Interpolation of operations

\[ f, \frac{df}{dt}, \frac{d^2f}{dt^2}, \frac{d^3f}{dt^3}, \ldots \]

\[ f, \int f(t) dt, \int dt \int f(t) dt, \int dt \int dt \int f(t) dt, \ldots \]

\[ \ldots, \frac{d^2f}{dt^2}, \frac{d^3f}{dt^3}, \frac{df}{dt}, \frac{df}{dt^2}, \ldots \]

---

**Fractional Calculus** was born in 1695

What if the order will be \( n = \frac{1}{2} \)?

It will lead to a paradox, from which one day useful consequences will be drawn.

G.F.A. de L'Hôpital (1661–1704)

G.W. Leibniz (1646–1716)

---

The beginning of a new stage

<table>
<thead>
<tr>
<th>1695</th>
<th>1960s</th>
<th>You are here</th>
</tr>
</thead>
<tbody>
<tr>
<td>static models</td>
<td>dynamical models</td>
<td>fractional order modeling</td>
</tr>
<tr>
<td>geometry, algebra</td>
<td>differential and integral calculus</td>
<td>Do better than fractional calculus</td>
</tr>
</tbody>
</table>

Do better than the best doable before!
Soft matters, complex fluids

Why BD and ML must meet FC?
“Fractional Order Thinking”
or, “In Between Thinking”

- For example
  - Between integers there are non-integers;
  - Between logic 0 and logic 1, there is the “fuzzy logic”;
  - Between integer order splines, there are “fractional order splines”
  - Between integer high order moments, there are noninteger order moments (e.g. FLOS)
  - Between “integer dimensions”, there are fractal dimensions
  - Fractional Fourier transform (FrFT) – in-between time-n-freq.
  - Non-Integer order calculus (fractional order calculus – abuse of terminology.) (FOC)
Rule of thumb for “Fractional Order Thinking”

- Self-similar
- Scale-free/Scale-invariant
- Power law
- Long range dependence (LRD)
- $1/f^\alpha$ noise
- Porous media
- Particulate
- Granular
- Lossy
- Anomaly
- Disorder
- Soil, tissue, electrodes, bio, nano, network, transport, diffusion, soft matters (biox) …
What is considered as complex?

（Inverse）Power Law （IPL）

https://www.zhihu.com/question/20313934

为什么我国的概率与统计学教科书里不怎么讲幂律分布？
Complexity “bow tie”  
(When you start to call it complex?)

Complex systems, phenomena, behaviors, …

Scale-Free, Heavy-Tailedness, Long Range Dependence, Long Memory …

Why BD and ML must meet FC?
### Empirical Power Laws

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Law’s name</th>
<th>Form of law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropology</td>
<td>Auerbach</td>
<td>$Pr(\text{city size rank } r) \propto 1/r$</td>
</tr>
<tr>
<td></td>
<td>War</td>
<td>$Pr(\text{intensity } &gt; I) \propto 1/I^{1/\alpha}$</td>
</tr>
<tr>
<td></td>
<td>1/f Music</td>
<td>Spectrum$(f) \propto 1/f$</td>
</tr>
<tr>
<td>Biology</td>
<td>DNA sequence</td>
<td>Symbol spectrum$(\text{frequency } f) \propto 1/f^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Ecological web</td>
<td>$Pr(\text{species connections}) \propto 1/k^{1.1}$</td>
</tr>
<tr>
<td></td>
<td>Protein</td>
<td>$Pr(\text{connections}) \propto 1/k^{2.4}$</td>
</tr>
<tr>
<td></td>
<td>Metabolism</td>
<td>$Pr(\text{connections}) \propto 1/k^{2.2}$</td>
</tr>
<tr>
<td></td>
<td>Sexual relations</td>
<td>$Pr(\text{relations}) \propto 1/k^{\alpha}$</td>
</tr>
<tr>
<td>Botany</td>
<td>da Vinci</td>
<td>Branching: $d_0^6 = d_1^2 + d_2^2$</td>
</tr>
<tr>
<td></td>
<td>Willis</td>
<td>No. of genera$(\text{No. of species } N) \propto 1/N^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Murray</td>
<td>$d_0^{2.5} = d_1^{2.5} + d_2^{2.5}$</td>
</tr>
<tr>
<td>Economics</td>
<td>Pareto</td>
<td>$Pr(\text{income } x) \propto 1/x^{1.5}$</td>
</tr>
<tr>
<td></td>
<td>Price variations</td>
<td>$Pr(\text{stock price variations}) \propto 1/x^{3}$</td>
</tr>
<tr>
<td>Geophysics</td>
<td>Omori</td>
<td>$Pr(\text{aftershocks in time } t) \propto 1/t$</td>
</tr>
<tr>
<td></td>
<td>Rosen–Rammel</td>
<td>$Pr(\text{No. of ore fragments } &lt; \text{size } r) \propto r^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Korčak</td>
<td>$Pr(\text{island area } A &gt; a) \propto 1/d^{2}$</td>
</tr>
<tr>
<td></td>
<td>Horton</td>
<td>$Pr(\text{segments at } n/No. \text{ of segments at } n + 1) \propto 1$</td>
</tr>
<tr>
<td></td>
<td>Gutenberg–Richter</td>
<td>$Pr(\text{earthquake magnitude } &lt; x) \propto 1/x^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Hack</td>
<td>$\text{River length } \propto (\text{basin area})^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Richardson</td>
<td>$\text{Length of coastline } \propto 1/(\text{ruler size})^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>Forest fires</td>
<td>$\text{Frequency density(burned area) } \propto 1/A^{1.38}$</td>
</tr>
<tr>
<td>Information theory</td>
<td>World Wide Web</td>
<td>$Pr(\text{connections}) \propto 1/k^{1.94}$</td>
</tr>
<tr>
<td></td>
<td>Internet</td>
<td>$Pr(\text{connections}) \propto 1/k^{\alpha}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Law’s name</th>
<th>Form of law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>1918 [70] 1/f noise</td>
<td>Spectrum$(f) \propto 1/f$</td>
</tr>
<tr>
<td></td>
<td>2002 [25] Solar flares</td>
<td>$Pr(\text{time between flares } t) \propto 1/t^{2.14}$</td>
</tr>
<tr>
<td></td>
<td>2003 [69] Temperature anomalies</td>
<td>$Pr(\text{time between events } t) \propto 1/t^{2.14}$</td>
</tr>
<tr>
<td>Physiology</td>
<td>1939 [61] Rall</td>
<td>$Pr(\text{isotope expelled in time } t) \propto 1/t^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>1963 [76] Mammalian vascular network</td>
<td>$\text{Veins and arteries } d_0^{1.5} = d_1^{1.5} + d_2^{1.5}$</td>
</tr>
<tr>
<td></td>
<td>1963 [90] Bronchial tree</td>
<td>$d_3^{2.7} = d_1^{2.7} + d_2^{2.7}$</td>
</tr>
<tr>
<td></td>
<td>1973 [48] McMahon</td>
<td>$d_0^{3} = d_1^{3} + d_2^{3}$</td>
</tr>
<tr>
<td></td>
<td>1976 [103] Radioactive clearance</td>
<td>$\text{Metabolic rate(body mass } M) \propto M^{0.75}$</td>
</tr>
<tr>
<td></td>
<td>1987 [93] West–Goldberger</td>
<td>$Pr(\text{heartbeat variability}) \propto f$</td>
</tr>
<tr>
<td></td>
<td>1991 [30] Mammalian brain</td>
<td>$\text{Power spectrum(frequency } f) \propto f$</td>
</tr>
<tr>
<td></td>
<td>1992 [77] Interbreath variability</td>
<td>$Pr(\text{time between EEG events}) \propto 1/t^{1.61}$</td>
</tr>
<tr>
<td></td>
<td>1993 [58] Heartbeat variability</td>
<td>$Pr(\text{behavior connections}) \propto 1/k^{\alpha}$</td>
</tr>
<tr>
<td>Psychology</td>
<td>1957 [75] Psychophysiology</td>
<td>Perceived response$(\text{stimulus intensity } x) \propto x^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>1963 [71] Trial and error</td>
<td>Reaction time$(\text{trial } N) \propto 1/N^{0.91}$</td>
</tr>
<tr>
<td></td>
<td>1961 [29] Decision making</td>
<td>utility$(\text{delay time } t) \propto t^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>1991 [3] Forgetting</td>
<td>Percentage correct recall$(\text{time } t) \propto 1/t^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>2001 [20] Cognition</td>
<td>Response spectrum$(\text{frequency } f) \propto 1/f^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>2009 [37] Neurophysiology</td>
<td>$Pr(\text{phase-locked interval } &lt; \tau) \propto 1/\tau^{\alpha}$</td>
</tr>
<tr>
<td>Sociology</td>
<td>1926 [41] Lotka</td>
<td>$Pr(\text{No. of papers published rank } r) \propto 1/r^{2}$</td>
</tr>
<tr>
<td></td>
<td>1949 [104] Zipf</td>
<td>$Pr(\text{word has rank } r) \propto 1/r$</td>
</tr>
<tr>
<td></td>
<td>1963 [16] Price</td>
<td>$Pr(\text{citation rank } r) \propto 1/r^{3}$</td>
</tr>
<tr>
<td></td>
<td>1994 [8] Urban growth</td>
<td>$\text{Population density(radius } R) \propto R^{\alpha}$</td>
</tr>
<tr>
<td></td>
<td>1998 [88] Actors</td>
<td>$Pr(\text{connections}) \propto 1/k^{2.3}$</td>
</tr>
</tbody>
</table>

---

IPL in Different Contexts

- Scale-free networks (degree distributions)
- Pink noise (power spectrum)
- Probability density function (PDF)
- Autocorrelation function (ACF)
- Allometry \((Y=a \; X^b)\)
- Anomalous relaxation (evolving over time)
- Anomalous diffusion (MSD versus time)
- Self-similar
Other connectedness to FC? (hidden)

- Fractal, irregular, anomalous, rough, Hurst
  - Multifractal, multi-scale, scale-rich
- Renormalization (?), Universality
- Extreme events—spikiness, bursty, intermittence
- Fluctuation in fluctuations; Variability,
- Emergence, Surprise, **Black swan**
- Nonlocality, Long term memory
- Complex (behavior, processes, network, fluid, dynamics, systems …)
- When the forest is big, there are all types of birds ("It takes all kinds" 林子大了什么鸟都有), 20/80 rule(二八定律)
My submission:

Fractional dynamics point of view of complex systems for complexity characterization and regulation
Why BD and ML must meet FC?
Fractional Calculus
View of Complexity
Tomorrow's Science
Bruce J. West
Bruce J. West has been a research scientist and teacher for forty years. He is one of a handful of scientists in the world that understands complexity and who can explain its implications for modern society in everyday language.

In *Complex Worlds: Uncertain, Unequal and Unfair* he uses his understanding of complex networks to explain why the future cannot be made certain, why the same people are always at the center of controversy, and why only a select few get ahead. The emerging properties of complexity so prevalent in society stand in sharp contrast to how the greatest thinkers of the past and present believe the world ought to be.

West explores the question: Is the dissonance between what is true and what we believe ought to be true really that great? The answer is a resounding yes and he explains not only how but why.

Dr. Bruce J. West, Ph.D., FAPS, FARL has had three careers. The first was as an Industry Researcher in a small not-for-profit The La Jolla Institute, 1971-1989. The second was as a Full Professor and Physics Department Chair at the University of North Texas, 1989-1999. The third is as Chief Scientist of Mathematics for the U.S. Army Research Office, 1999-present.

Why BD and ML must meet FC?
Power Law

\[ f(x) = ax^k \]

When \( k \) is negative: \textit{Inverse power law}

\[ f(cx) = a(cx)^k = c^k f(x) \propto f(x). \]

Why BD and ML must meet FC?

- Fractional dynamic
- Mittag-Leffler Law
- Fractional calculus based models
- Scale-Free, Heavy-Tailedness, Long Range Dependence, Long Memory ...
- Complex systems, phenomena, behaviors, ...

07/06/2020
Root of long (algebraic) tail, or inverse power law

\[ E_{\alpha,1}(-x^\alpha) \approx \begin{cases} 
1 - x^\alpha / \Gamma(1 + \alpha) & \text{as } x \to 0^+ \\
x^{-\alpha} / \Gamma(1 - \alpha) & \text{as } x \to \infty 
\end{cases} \]
Heavy tail, fat tail

\[ P[X > x] \sim x^{-\alpha} \]
Brownian motion

Lévy flights

Wandering albatrosses flight search patterns

Complex signals

- Long-range dependence
- Hurst parameter
- ARFIMA
- Fractional Gaussian noise (FGn)
- Self-similar
- \(\alpha\)-stable distributions
- "Spikiness" intermittence
- "Heavy tails"
- Fractional Brownian motion (FBm)

- Network Traffic, Smart Grids
- HRV, Outliers in time series, biox signals ...
Connection to FC via PDF


\[
\frac{\partial u}{\partial t} = D(\alpha) \frac{\partial^\alpha u}{\partial |x|^\alpha}, \quad -\infty < x < +\infty, \quad t \geq 0,
\]

with \( u(x,0) = \delta(x) \quad 0 < \alpha \leq 2 \)

\[
\frac{\partial^{2\beta} u}{\partial t^{2\beta}} = D(\beta) \frac{\partial^2 u}{\partial x^2}, \quad x \geq 0, \quad t \geq 0,
\]

with \( u(0,t) = \delta(t) \quad 0 < \beta < 1 \)

07/06/2020

Why BD and ML must meet FC?
• Can these be synthesized?

• Is the fractional calculus entailed by complexity?
Take home messages

• Triangle:
  • Complexity
  • Power Law
  • Fractional Calculus

• Stochasticity with rich forms
  (heavytailedness)
  • Fractional order master equations
Outline

• Fractional Calculus, Complexity, and Fractional Order Thinking

• Big Data, Variability, and Fractional Calculus

• Machine Learning, Optimal Randomness and Fractional Calculus

• Looking Into Future: Fractional Calculus is Physics Informed
10 V’s of Big Data

- #1: Volume
- #2: Velocity
- #3: Variety
- **#4: Variability**
- #5: Veracity
- #6: Validity
- #7: Vulnerability
- #8: Volatility
- #9: Visualization
- #10: Value

https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
When talking about big data, we have to talk about “10V”s.

- Variability in big data's context refers to a few different things. One is the number of inconsistencies in the data. These need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur.
- Variability can also refer to diversity. In practice, the data can be classified into several different types, for example, healthy or unhealthy.

https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
#1: Volume

- 300 hours of video are uploaded to YouTube every minute.
- An estimated 1.1 trillion photos were taken in 2016, and that number is projected to rise by 9 percent in 2017. As the same photo usually has multiple instances stored across different devices, photo or document sharing services as well as social media services, the total number of photos stored is also expected to grow from 3.9 trillion in 2016 to 4.7 trillion in 2017.
- In 2016 estimated global mobile traffic amounted for 6.2 exabytes per month. That's 6.2 billion gigabytes.
#2: Velocity

- Velocity refers to the speed at which data is being generated, produced, created, or refreshed.
- Sure, it sounds impressive that Facebook's data warehouse stores upwards of 300 petabytes of data, but the velocity at which new data is created should be taken into account. Facebook claims 600 terabytes of incoming data per day.
- Google alone processes on average more than "40,000 search queries every second," which roughly translates to more than 3.5 billion searches per day.
#3: Variety 多样化

- When it comes to big data, we don't only have to handle structured data but also semistructured and mostly unstructured data as well. As you can deduce from the above examples, most big data seems to be unstructured, but besides audio, image, video files, social media updates, and other text formats there are also log files, click data, machine and sensor data, etc.
#4: Variability 变化性

- Variability in big data's context refers to a few different things. One is the number of inconsistencies in the data. These need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur.

- Big data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources. Variability can also refer to the inconsistent speed at which big data is loaded into your database.
#5: Veracity 真実性

• This is one of the unfortunate characteristics of big data. As any or all of the above properties increase, the veracity (confidence or trust in the data) drops. This is similar to, but not the same as, validity or volatility (see below). Veracity refers more to the provenance or reliability of the data source, its context, and how meaningful it is to the analysis based on it.
#6: Validity 有效性

- Similar to veracity, validity refers to how accurate and correct the data is for its intended use. According to Forbes, an estimated 60 percent of a data scientist's time is spent cleansing their data before being able to do any analysis. The benefit from big data analytics is only as good as its underlying data, so you need to adopt good data governance practices to ensure consistent data quality, common definitions, and metadata.
Big data brings new security concerns. After all, a data breach with big data is a big breach. Does anyone remember the infamous AshleyMadison hack in 2015?

Unfortunately there have been many big data breaches. Another example, as reported by CRN: in May 2016 "a hacker called Peace posted data on the dark web to sell, which allegedly included information on 167 million LinkedIn accounts and ... 360 million emails and passwords for MySpace users."

Information on many others can be found at Information is Beautiful.
#8: Volatility 挥发性

- How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?

- Before big data, organizations tended to store data indefinitely -- a few terabytes of data might not create high storage expenses; it could even be kept in the live database without causing performance issues. In a classical data setting, there not might even be data archival policies in place.

- Due to the velocity and volume of big data, however, its volatility needs to be carefully considered. You now need to establish rules for data currency and availability as well as ensure rapid retrieval of information when required. Make sure these are clearly tied to your business needs and processes -- with big data the costs and complexity of a storage and retrieval process are magnified.
#9: Visualization

- Current big data visualization tools face technical challenges due to limitations of in-memory technology and poor scalability, functionality, and response time. You can't rely on traditional graphs when trying to plot a billion data points, so you need different ways of representing data such as data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees.
- Combine this with the multitude of variables resulting from big data's variety and velocity and the complex relationships between them, and you can see that developing a meaningful visualization is not easy.
#10: Value

- Last, but arguably the most important of all, is value. The other characteristics of big data are meaningless if you don't derive business value from the data.

- Substantial value can be found in big data, including understanding your customers better, targeting them accordingly, optimizing processes, and improving machine or business performance. You need to understand the potential, along with the more challenging characteristics, before embarking on a big data strategy.
万变不离其宗

荀子·儒效. 《荀子》是战国时期(475-221BC)荀子和弟子们整理或记录他人言行的哲学著作。

What is “宗” for Big Data?

- Variability

“The Only Thing That Is Constant Is Change” — Heraclitus 赫拉克利特(纪元前五世纪的希腊哲学家)

07/06/2020 Why BD and ML must meet FC?
(Fractional Order) Dynamics: per pre-Socratic Greek philosopher, Heraclitus

• “It is in changing that we find purpose.”
  — Heraclitus

• “Nothing endures but change.”
  — Heraclitus

• “No man ever steps in the same river twice, for it's not the same river and he's not the same man.”
  — Heraclitus
2000+ years later
Integer Order Calculus

- Invented late 17th century by Isaac Newton and Gottfried Wilhelm Leibniz
- Note: Integer Order Dynamic view of Changes is only for our own “convenience” (Scott Blair)
- Denying fractional calculus is as saying there is no non-integers in between integers!
“We may express our concepts in Newtonian terms if we find this convenient but, if we do so, we must realize that we have made a translation into a language which is foreign to the system which we are studying.” (1950)
Why BD and ML must meet FC?
(anything’s) variability

- **Climate variability**, changes in the components of Earth's climate system and their interactions
- **Genetic variability**, a measure of the tendency of individual genotypes in a population to vary from one another
- **Heart rate variability**, a physiological phenomenon where the time interval between heart beats varies
- **Human variability**, the range of possible values for any measurable characteristic, physical or mental, of human beings
- **Spatial variability**, when a quantity that is measured at different spatial locations exhibits values that differ across the locations
- **Statistical variability**, a measure of dispersion in statistics

- Gait, breath, cognitive, temperature, soil, crop, ……

Cosmic Background Radiation
https://i.ytimg.com/vi/WB5jmdJvQeU/maxresdefault.jpg
07/06/2020

Why BD and ML must meet FC?
Variability in universe? True


Why BD and ML must meet FC?
Variability in “big data”?

• Sure!
• But how MAD about it?
  – We need Fractional Calculus!

So, to be complex to have big data??

• Sure!
• But how MAD about it?
  – We need Fractional Calculus!
Yes, BD should meet FC!

HOW?
**FODA**: Fractional Order Data Analytics

- First proposed by Prof. YangQuan Chen 2015.
- Metrics based on using fractional order signal processing techniques for quantifying the generating dynamics of observed or perceived variabilities.
  - Hurst parameter, fGn, fBm, …
  - Fractional order integral, differentiation
  - FLOM/FLOS (fractional order lower order moments/statistics)
  - Alpha stable processes, Levy flights
  - ARFIMA, GARMA (Gegenbauer), CTRW
Fractional Order Data Analytics: connecting dots of Drones, Big Data, and Fractional Calculus

YangQuan Chen, Ph.D., Director,
ME/EECS/SNRI/HSRI, School of Engineering,
University of California, Merced
E: yqchen@ieee.org; or, yangquan.chen@ucmerced.edu
T: (209)228-4672; O: SE1-254; Lab: CAS Eng 820 (T: 228-4398)

March 21, 2015. Saturday 2:00-2:15 PM
Robots & Ribs Day @ MESA LAB Symposium @ UC Merced
Why BD and ML must meet FC?

https://www.exelisvis.com/Learn/WhitepapersDetail/TabId/802/ArtMID/2627/ArticleID/13742/Vegetation-Analysis-Using-Vegetation-Indices-in-ENVI.aspx
Decreased photosynthetic response and pigmentation

Increased secondary metabolite biosynthesis

Increased secondary metabolite biosynthesis

Osmoprotection

Root development

Signal transduction (ethylene-ABA-JA)

Stomatal closure

Light energy

CO₂

O₂

Complex

Why BD and ML must meet FC?
NDVI vs. water stress??

Graphs showing the relationship between NDVI and SWP (Soil Water Potential) with and without a shaded region. The equations and correlation coefficients are indicated for each graph:

- **NDVI vs. SWP (with shaded region):**
  - Equation: \( y = -0.0008x + 0.4407 \)
  - \( R^2 = 0.001 \)

- **NDVI vs. SWP (without shaded region):**
  - Equation: \( y = -0.0012x + 0.4559 \)
  - \( R^2 = 0.0052 \)
Drones as “Tractor 2.0” for Farmers

• RRR or SSM of water, fertilizers, pesticides etc.
• Fractional Calculus may save the world one day.
• Drones create big data and demand FODA due to “complexity” thus variability, inherent in life process.
Outline

• Fractional Calculus, Complexity, and Fractional Order Thinking
• Big Data, Variability, and Fractional Calculus
• Machine Learning, Optimal Randomness and Fractional Calculus
• Looking Into Future: Fractional Calculus is Physics Informed
Why BD and ML must meet FC?
MESA LAB Learning Data Optimization methods Speed Accuracy Convergence rate Optimization methods Global Searching More Intelligent More 天人合一

人，社会，和自然 vs. 战天斗地 人定胜天？

More 天人合一
学习（算法）很重要，不然会累死人的。

Learning algorithm is important, or we will tire to death.
Reflection 沉思：

- (Machine) Learning is now a hot research topic;
- How to learn efficiently (optimally) is always important;
- The key for learning is the optimization method;
- Thus designing an efficient optimization method is the most important topic now
  - What is the optimal way to optimize?
  - What is the more optimal way to optimize?

ML core is optimization, can we demand “More Optimal Machine Learning” (i.e., DL with minimum/smallest labelled data)?
Levy flight is *optimized* randomness for albatrosses via millions of years of evolution or slow optimization.

Levy flight is optimized randomness for albatrosses via millions of years of evolution or slow optimization.

Can it be more optimal?

Wandering albatrosses

Nowadays, Big Data and Machine Learning are two hottest topics and they are closely related to each other. To better understand them, we also need F.C. （分数阶微积分） and H.T. （重尾）
https://math.stackexchange.com/questions/3664716/taxonomy-overview-of-optimization-methods

07/06/2020

Why BD and ML must meet FC?
Two broad categories

- **Derivative-free**
  - Direct Search, NM, PSO

- **Gradient-based**
  - GD and its variants
Derivative-free

Single agent search vs. swarm-based search

2-D Alpine function

\[ f(x) = f(x_1, \ldots, x_n) = \sum_{i=1}^{n} |x_i \sin(x_i) + 0.1x_i | \]
Exploration is often achieved by randomization or random numbers in terms of some predefined probability distributions.

Exploitation uses local information such as gradients to search local regions more intensively, and such intensification can enhance the rate of convergence.

What is the optimal randomness?

J Wei, YQ Chen, Y Yu, Y Chen (2019). Optimal Randomness in Swarm-Based Search. Mathematics 7 (9), 828 [PDF]
4 HT 重尾 distributions – sample paths

- Mitig - Lefler distribution
- Weibull distribution
- Pareto distribution
- Cauchy distribution

Why BD and ML must meet FC?
### 2.3 Experimental results

<table>
<thead>
<tr>
<th>No.</th>
<th>Test function</th>
<th>Parameters configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$F_{sph}$: Sphere’s Function</td>
<td>$NP = 20$ (population size),</td>
</tr>
<tr>
<td>2</td>
<td>$F_{ros}$: Rosenbrock’s Function</td>
<td>$P_a = 0.25$ (discovery probability),</td>
</tr>
<tr>
<td>3</td>
<td>$F_{ack}$: Ackley’s Function</td>
<td>$\text{Max}_\text{FEs}=10,000*D$ (termination criterion),</td>
</tr>
<tr>
<td>4</td>
<td>$F_{grw}$: Griewank’s Function</td>
<td>$\text{runs} = 50$ (Running times)</td>
</tr>
<tr>
<td>5</td>
<td>$F_{ras}$: Rastrigin’s Function</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$F_{sch}$: Generalized Schwefel’s Problem 2.26</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$F_{sal}$: Salomon’s Function</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$F_{whl}$: Whitely’s Function</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$F_{pn1}$: Generalized Penalized Function 1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$F_{pn2}$: Generalized Penalized Function 2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>$F_1$: Shifted Sphere Function</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>$F_2$: Shifted Schwefel’s Problem 1.2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>$F_3$: Shifted Rotated High Conditioned Elliptic Function</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>$F_4$: Shifted Schwefel’s Problem 1.2 with Noise in Fitness</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>$F_5$: Schwefel’s Problem 2.6 with global Optimum on Bounds</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>$F_6$: Shifted Rosenbrock’s Function</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>$F_7$: Shifted Rotated Griewank’s Function without Bounds</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>$F_8$: Shifted Rotated Ackley’s Function with Global Optimum on Bounds</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>$F_9$: Shifted Rastrigin’s Function</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>$F_{10}$: Shifted Rotated Rastrigin’s Function</td>
<td></td>
</tr>
</tbody>
</table>

07/06/2020

Why BD and ML must meet FC?
<table>
<thead>
<tr>
<th>Fun</th>
<th>CS</th>
<th>CSML</th>
<th>CSP</th>
<th>CSC</th>
<th>CSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{sph}$</td>
<td>9.58E-31</td>
<td>4.90E-54‡</td>
<td>4.74E-59‡</td>
<td>1.17E-57‡</td>
<td>4.40E-51‡</td>
</tr>
<tr>
<td>$F_{ros}$</td>
<td>1.20E+01</td>
<td>5.22E+00‡</td>
<td>3.10E+00‡</td>
<td>2.74E+00‡</td>
<td>8.62E+00‡</td>
</tr>
<tr>
<td>$F_{ack}$</td>
<td>7.70E-13</td>
<td>1.06E-14‡</td>
<td>1.07E-14‡</td>
<td>9.56E-15‡</td>
<td>8.28E-15‡</td>
</tr>
<tr>
<td>$F_{grw}$</td>
<td>7.11E-17</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
</tr>
<tr>
<td>$F_{ras}$</td>
<td>2.32E+01</td>
<td>1.38E+01‡</td>
<td>1.88E+01‡</td>
<td>1.49E+01‡</td>
<td>8.34E+00‡</td>
</tr>
<tr>
<td>$F_{sch}$</td>
<td>1.57E+03</td>
<td>5.37E+02‡</td>
<td>1.32E+03‡</td>
<td>4.80E+02‡</td>
<td>3.56E+01‡</td>
</tr>
<tr>
<td>$F_{sal}$</td>
<td>3.76E-01</td>
<td>2.96E-01‡</td>
<td>3.00E-01‡</td>
<td>2.84E-01‡</td>
<td>2.20E-01‡</td>
</tr>
<tr>
<td>$F_{whi}$</td>
<td>3.73E+02</td>
<td>2.00E+02‡</td>
<td>2.49E+02‡</td>
<td>2.27E+02‡</td>
<td>1.93E+02‡</td>
</tr>
<tr>
<td>$F_{pc}$</td>
<td>2.07E-03</td>
<td>1.57E-32‡</td>
<td>1.57E-32‡</td>
<td>2.07E-03$^\infty$</td>
<td>1.57E-32‡</td>
</tr>
<tr>
<td>$F_{pn1}$</td>
<td>4.82E-28</td>
<td>1.35E-32‡</td>
<td>1.35E-32‡</td>
<td>1.35E-32‡</td>
<td>1.35E-32‡</td>
</tr>
<tr>
<td>$F_{pn2}$</td>
<td>6.48E-30</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
<td>0.00E+00*</td>
</tr>
<tr>
<td>$F_{2}$</td>
<td>1.05E-02</td>
<td>1.10E-03‡</td>
<td>2.77E-04‡</td>
<td>1.40E-03‡</td>
<td>1.23E-02‡</td>
</tr>
<tr>
<td>$F_{3}$</td>
<td>2.17E+06</td>
<td>3.04E+06†</td>
<td>2.99E+06†</td>
<td>3.25E+06†</td>
<td>3.61E+06†</td>
</tr>
<tr>
<td>$F_{4}$</td>
<td>1.79E+03</td>
<td>4.98E+02‡</td>
<td>3.58E+02‡</td>
<td>4.02E+02‡</td>
<td>5.51E+02‡</td>
</tr>
<tr>
<td>$F_{5}$</td>
<td>3.17E+03</td>
<td>2.44E+03‡</td>
<td>1.98E+03‡</td>
<td>2.11E+03‡</td>
<td>1.94E+03‡</td>
</tr>
<tr>
<td>$F_{6}$</td>
<td>2.78E+01</td>
<td>1.57E+01‡</td>
<td>9.91E+00‡</td>
<td>1.23E+01‡</td>
<td>1.59E+01‡</td>
</tr>
<tr>
<td>$F_{7}$</td>
<td>1.34E-03</td>
<td>2.22E-03†</td>
<td>5.79E-03†</td>
<td>3.73E-03†</td>
<td>2.49E-03†</td>
</tr>
<tr>
<td>$F_{8}$</td>
<td>2.09E+01</td>
<td>2.09E+01$^\infty$</td>
<td>2.09E+01$^\infty$</td>
<td>2.09E+01$^\infty$</td>
<td>2.09E+01$^\infty$</td>
</tr>
<tr>
<td>$F_{9}$</td>
<td>2.84E+01</td>
<td>1.30E+01‡</td>
<td>2.74E+01‡</td>
<td>1.28E+01‡</td>
<td>6.81E+00‡</td>
</tr>
<tr>
<td>$F_{10}$</td>
<td>1.69E+02</td>
<td>1.21E+02‡</td>
<td>1.31E+02‡</td>
<td>1.18E+02‡</td>
<td>1.03E+02‡</td>
</tr>
<tr>
<td>$\div/\approx/\dagger$</td>
<td>-</td>
<td>17/1/2</td>
<td>17/1/2</td>
<td>16/2/2</td>
<td>16/1/3</td>
</tr>
<tr>
<td>p-value</td>
<td>8.97E-03</td>
<td>1.00E-02</td>
<td>1.00E-02</td>
<td>1.87E-02</td>
<td></td>
</tr>
<tr>
<td>Avg. rank</td>
<td>4.35</td>
<td>2.78</td>
<td>2.88</td>
<td>2.58</td>
<td>2.43</td>
</tr>
</tbody>
</table>
Figure 2. Convergence curves of CS and different improved CS algorithms for selected functions at $D = 30$. Why BD and ML must meet FC?
Connection to FC via PDF


\[
\frac{\partial u}{\partial t} = D(\alpha) \frac{\partial^{\alpha} u}{\partial |x|^{\alpha}}, \quad -\infty < x < +\infty, \quad t \geq 0, \\
\text{with} \quad u(x, 0) = \delta(x) \quad 0 < \alpha \leq 2
\]

\[
\frac{\partial^{2\beta} u}{\partial t^{2\beta}} = D(\beta) \frac{\partial^{2} u}{\partial x^{2}}, \quad x \geq 0, \quad t \geq 0, \\
\text{with} \quad u(0, t) = \delta(t) \quad 0 < \beta < 1
\]
Optimal randomness means fractional calculus!

More optimal than the “optimized PSO” CS (Cuckoo Search)

07/06/2020

Why BD and ML must meet FC?
Two broad categories

- Derivative-free (fractional calculus helps)
  - Direct Search, NM, PSO
- Gradient-based
  - GD and its variants

07/06/2020 Why BD and ML must meet FC?
GD and SGD

Task: \( \min_x f(x) \)

GD does iterative updates \( x_{t+1} = x_t - \eta_t \nabla f(x_t) \)

\[ x_{t+1} = x_t - \eta_t g_t, \text{ where } \mathbb{E}[g_t] = \nabla f(x_t) \]

\[ g_t = \nabla f(x_t) + \xi_t \]

\( \xi_t \in \mathcal{N}(0, I) \text{ or } \mathbb{B}_0(r) \)

Why BD and ML must meet FC?
MESA LAB

- Momentum GD
- Accelerated GD
- Nesterov GD
- Noisy GD
- Stochastic GD
- Perturbed GD
- Other Variants

NAGD
**Why BD and ML must meet FC?**

---

<table>
<thead>
<tr>
<th>Title</th>
<th>Cited by</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior-point polynomial algorithms in convex programming</td>
<td>5426</td>
<td>1994</td>
</tr>
<tr>
<td>Introductory lectures on convex optimization: A basic course</td>
<td>3522</td>
<td>2013</td>
</tr>
<tr>
<td>A method for solving the convex programming problem with convergence rate $O(1/k^2)$</td>
<td>2495*</td>
<td>1983</td>
</tr>
<tr>
<td>Smooth minimization of non-smooth functions</td>
<td>1997</td>
<td>2005</td>
</tr>
<tr>
<td>Gradient methods for minimizing composite objective function</td>
<td>1577*</td>
<td>2007</td>
</tr>
</tbody>
</table>
NAGD: Nesterov Accelerated GD

\[
\begin{align*}
    y_{k+1} &= a y_k - \mu \nabla f(x_k) \\
    x_{k+1} &= x_k + y_{k+1} + b y_k
\end{align*}
\]

- Set \( b = \frac{-a}{1+a} \) and one can derive the NAGD.

- Set \( b=0 \) and one can derive GDM.


07/06/2020 Why BD and ML must meet FC?
Michael I. Jordan: Is there an optimal way to optimize?

- ICM2018 1 hour report “Dynamical, symplectic and stochastic perspectives on optimization” [https://youtu.be/wXNWVhE2Dl4](https://youtu.be/wXNWVhE2Dl4)
Why BD and ML must meet FC?

<table>
<thead>
<tr>
<th>Title</th>
<th>Cited by</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent dirichlet allocation</td>
<td>25421</td>
<td>2003</td>
</tr>
<tr>
<td>DM Blei, AY Ng, MI Jordan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of machine Learning research 3 (Jan), 993-1022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On spectral clustering: Analysis and an algorithm</td>
<td>7142</td>
<td>2002</td>
</tr>
<tr>
<td>AY Ng, MI Jordan, Y Weiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advances in neural information processing systems, 849-856</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive mixtures of local experts</td>
<td>3841</td>
<td>1991</td>
</tr>
<tr>
<td>RA Jacobs, MI Jordan, SJ Nowlan, GE Hinton</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural computation 3 (1), 79-87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharing clusters among related groups: Hierarchical Dirichlet processes</td>
<td>3607</td>
<td>2005</td>
</tr>
<tr>
<td>YW Teh, MI Jordan, MJ Beal, DM Blei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advances in neural information processing systems, 1385-1392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical mixtures of experts and the EM algorithm</td>
<td>3202</td>
<td>1994</td>
</tr>
<tr>
<td>MI Jordan, RA Jacobs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural computation 6 (2), 481-514</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Review of Michael Jordan’s work:

Nesterov accelerated GD (NAGD) can be formulated as

\[
\begin{align*}
    x_k &= y_{k-1} - \mu \nabla f(y_{k-1}) \\
    y_k &= x_k + \frac{k-1}{k+2} (x_k - x_{k-1})
\end{align*}
\]

Set \( t = k\sqrt{\mu} \) and one can derive its corresponding differential equation

\[
\ddot{X} + \frac{3}{t} \dot{X} + \nabla f(X) = 0
\]

- The main idea of Michael Jordan’s work is to analyze the iteration algorithm in the continuous-time domain.
- For differential equation, one can use Lyapunov method or variational method to analyze its properties.
Review of Michael Jordan’s work

Take Lyapunov functional as

\[ V(t) = t^2 (f(X(t)) - f^*) + 2 \left\| X + \frac{t}{2} \dot{X} - x^* \right\| \]

whose time derivative is

\[ \dot{V}(t) = 2t (f(X(t)) - f^*) - 2t \langle X - x^*, \nabla f(X) \rangle \leq 0 \]

due to the convexity of \( f(X) \).

Then one has

\[ f(X(t)) - f^* \leq \frac{V(t)}{t^2} \leq \frac{V(0)}{t^2} = \frac{2\|x_0 - x^*\|^2}{t^2} \]

which indicates a \( O\left(\frac{1}{t^2}\right) \) convergence rate.
One can also use the variational method to derive the master differential equation for an optimization method.

**Variational principle**

- **Maupertuis**: Least Action Principle
- **Hamilton**: Hamilton’s Variational Principle
- **Feynman**: Quantum-Mechanical Path Integral Approach

Pierre-Louis Moreau de Maupertuis (1698 – 1759)

Sir William Rowan Hamilton (1805 – 1865)

Richard Phillips Feynman (1918 – 1988)
Review of Michael Jordan’s work

Define the Lagrangian

\[ \mathcal{L} = \mathcal{L}[q_1, q_2, \ldots, q_N, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_N, t] \]

where \((q_1, q_2, \ldots, q_N)\) are generalized coordinates.

We will then make the action integral stationary

\[ S = \int \mathcal{L}[q_1, q_2, \ldots, q_N, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_N, t] \, dt, \]

which requires that \(\mathcal{L}\) satisfy

\[ \frac{\partial \mathcal{L}}{\partial q_1} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_1}; \quad \frac{\partial \mathcal{L}}{\partial q_2} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_2}; \ldots; \quad \frac{\partial \mathcal{L}}{\partial q_N} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_N}. \]

\[ Nesterov accelerated GD (NAGD) can be formulated as \]

\[ \begin{cases} 
  x_k = y_{k-1} - \mu \nabla f(y_{k-1}) \\
  y_k = x_k + \frac{\mu}{k+1} (x_k - x_{k-1})
\end{cases} \]

Set \(t = k\sqrt{\mu}\) and one can derive its corresponding differential equation

\[ \ddot{x} + \frac{3}{t} \dot{x} + \nabla f(x) = 0 \]
Review of Michael Jordan’s work

Consider the convex function $f(x)$ and define the Bregman divergence

$$D_h (y, x) = h (y) - h (x) - \langle \nabla h (x) , y - x \rangle$$

where, $h(x)$ is also a convex function.

Define the second Bregman Lagrangian as

$$L (x, v, t) = e^{\alpha_t + \beta_t + \gamma_t} \left( \mu D_h (x, x + e^{-\alpha_t} v) - f (x) \right)$$

with ideal scaling conditions

$$\dot{\gamma}_t = e^{\alpha_t}$$
$$\dot{\beta}_t \leq e^{\alpha_t}$$
Review of Michael Jordan’s work

The second Bregman Lagrangian

\[ L(x, v, t) = e^{\alpha_t + \beta_t + \gamma_t} (\mu D_h(x, x + e^{-\alpha_t} v) - f(x)) \]

- \( D_h(x, x + e^{-\alpha_t} v) \) can be viewed as the kinetic energy and \(-f(x)\) is the potential energy.

- \( \alpha_t, \beta_t, \gamma_t \) are arbitrary smooth functions to help analyzing the convergence rate.
Review of Michael Jordan’s work

Define a functional on curves via integration of the Lagrangian:

\[ J = \int_R L(x, v, t) dt \]

and the Euler-Lagrange equation is

\[ \frac{\partial}{\partial x} L(x, v, t) = \frac{d}{dt} \frac{\partial}{\partial v} L(x, v, t) \]

The Euler-Lagrange equation for the second Bregman Lagrangian reduces to

\[ \frac{d}{dt} \nabla h \left( X_t + e^{\alpha_t} \dot{X}_t \right) = \dot{\beta}_t \nabla h (X_t) - \dot{\beta}_t \nabla h \left( X_t + e^{\alpha_t} \dot{X}_t \right) - \frac{e^{\alpha_t}}{\mu} \nabla f (X_t) \]
Review of Michael Jordan’s work

Define $h = \frac{1}{2} \| x \|^2$ and $\beta_t = \gamma t$, and one has

$$\ddot{X}_t + 2\gamma \dot{X}_t + \frac{\gamma^2}{\mu} \nabla f (X_t) = 0$$

which in the same form as the master differential equation of NAGD.

Assume $f$ is $\mu$-uniformly convex with respect to $h$ (strictly convex) and the scaling condition $\dot{\beta}_t = e^{\alpha t}$ holds. One can conclude that

$$V(t) = e^{\beta_t} \left( \mu D_h \left( x, X_t + e^{-\alpha t} \dot{X}_t \right) + f(X_t) - f(x) \right)$$

is a Lyapunov functional ($\dot{V} \leq 0$), which indicates a $O(e^{-\beta_t})$ convergence rate.
Review of Michael Jordan’s work

Discretize the previous Lagrangian system, and we can get following two discrete algorithms:

\[
\begin{aligned}
    x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k \\
    \nabla h(z_{k+1}) &= \nabla h(z_k) - \alpha_k \nabla f(x_{k+1}) \\
    y_{k+1} &= \Theta(x)
\end{aligned}
\]

and

\[
\begin{aligned}
    x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k \\
    y_{k+1} &= \Theta(x) \\
    \nabla h(z_{k+1}) &= \nabla h(z_k) - \alpha_k \nabla f(y_{k+1})
\end{aligned}
\]

where, \( \tau_k = \frac{A_{k+1} - A_k}{A_k} : = \frac{\alpha_k}{A_k} \), and \( \frac{d}{dt} e^{\beta t} = (A_{k+1} - A_k) / \delta \) can be well approximated by a discrete-time sequence \( A_i \). Moreover, \( \Theta \) is an arbitrary map whose domain is the previous state \( x = (x_{k+1}, z_{k+1} / z_k, y_k) \)

Review of Michael Jordan’s work

- One can transform an iterative (optimization) algorithm to its continuous-time limit case, which can simplify the analyses (Lyapunov methods).

- One can directly design a differential equation (of motion) and then discretize it to derive an iterative algorithm (Variational method).

- The key is to find a suitable Lyapunov functional to analyze the stability and convergent rate.

**Exciting new Fact:** Optimization algorithms can be systematically synthesized using Lagrangian mechanics (E-L) EOM.
Obviously, why not fractional order?

Define \( h = \frac{1}{2} \|x\|^2 \) and \( \beta_t = \gamma t \), and one has

\[
\ddot{X}_t + 2\gamma \dot{X}_t + \frac{\gamma^2}{\mu} \nabla f (X_t) = 0
\]

which in the same form as the master differential equation of NAGD.

\[
\dot{x}(t) \Rightarrow x^{(\alpha)}(t)
\]

Could to be “More Optimal”? (Better than the best)

Go “fractional” to be more optimal!

Fractional calculus of variation (FOCV), FO EL equation
GD and SGD

Task: $\min_x f(x)$

GD does iterative updates $x_{t+1} = x_t - \eta_t \nabla f(x_t)$

$$x_{t+1} = x_t - \eta_t g_t,$$ where $\mathbb{E}[g_t] = \nabla f(x_t)$$

$$g_t = \nabla f(x_t) + \xi_t$$

$$\xi_t \in \mathcal{N}(0, I) \text{ or } \mathcal{B}_0(r)$$

Why BD and ML must meet FC?
• Optimal randomness using fractional order noises can offer better than the best performance, similarly shown in

J Wei, YQ Chen, Y Yu, Y Chen (2019). *Optimal Randomness in Swarm-Based Search*. Mathematics 7 (9), 828 [PDF]

What control community can/should offer to CS/ML community?

• “The Three Musketeers”
  – Internal model principle (IMP)
  – Nu-Gap metric
  – Model discrimination

Eric Kerrigan,
Imperial College London, UK

07/06/2020
Analyses & Design with System Theory

- In [1], the authors transfer the convergence problem of numerical algorithms into a stability problem of a discrete-time system;
- In [2], the authors explained that the commonly used SGD-Momentum algorithm in Machine Learning is a PI controller and designed a PID algorithm.
- Motivated by [2] and different from Michael Jordan’s work, we will directly design and analyze the algorithms in S or Z domain.

Gradient Descent (GD) is a first-order algorithm:

\[ x_{k+1} = x_k - \mu \nabla f(x_k) \]

where \( \mu > 0 \) is the step size.

Using the Z-transform, we have that

\[ X(z) = \frac{\mu}{z - 1} [-\nabla f(x_k)]_z \]
Approximate the gradient around the extreme point $x^*$, and one has $\nabla f(x_k) \approx A(x_k - x^*)$ with $A = \nabla^2 f(x^*)$.

For GD, we have that $G(z) = \frac{1}{z-1}$, which is an integrator.

- Integrator in the forward loop is to eliminate the tracking error for a constant reference signal (Internal Model Principle).
Analyses & Design with System Theory

*GD-Momentum (GDM)* is then designed to accelerate the conventional GD, which is popularly used in Machine Learning.

\[
\begin{align*}
    y_{k+1} &= \alpha y_k - \mu \nabla f(x_k) \\
    x_{k+1} &= x_k + y_{k+1}
\end{align*}
\]

where \( y_k \) is the accumulation of history gradient and \( \alpha \in (0, 1) \) is the rate of moving average decay.

Do Z-transform for the update rule and derive

\[
\begin{align*}
    zY(z) &= \alpha Y(z) - \mu [\nabla f(x_k)]_z \\
    zX(z) &= X(z) + zY(z)
\end{align*}
\]
Analyses & Design with System Theory

Then one has

\[ X(z) = \frac{\mu z}{(z - 1)(z - \alpha)} \left[ -\nabla f(x_k) \right]_z \]

For GD-Momentum, we have that \( G(z) = \frac{z}{(z-1)(z-\alpha)} \), with an integrator in the forward loop.

- **GD-Momentum** is a second-order \((G(z))\) algorithm with an additional pole \(\alpha\) and zero \(0\).
- The “second-order” means the order of \(G(z)\), which is different from the algorithm using the Hessian matrix information.
Analyses & Design with System Theory

NAGD can be simplified as

\[
\begin{align*}
y_{k+1} &= x_k - \mu \nabla f(x_k) \\
x_{k+1} &= (1 - \lambda) y_{k+1} + \lambda y_k
\end{align*}
\]

where \( \mu \) is the step size and \( \lambda \) is a weighting coefficient.

Do Z-transform for the update rule and derive

\[
\begin{align*}
zY(z) &= X(z) - \mu [\nabla f(x_k)]_z \\
zX(z) &= (1 - \lambda) zY(z) + \lambda Y(z)
\end{align*}
\]
Then one has

$$X(z) = \frac{-(1-\lambda)z-\lambda}{(z-1)(z+\lambda)} \mu [\nabla f(x_k)]_z$$

$$= \frac{z+\frac{1-\lambda}{1-\lambda}}{(z-1)(z+\lambda)} \mu (1-\lambda) [-\nabla f(x_k)]_z$$

For NAGD, we have that $G(z) = \frac{z+\frac{1}{1-\lambda}}{(z-1)(z+\lambda)}$, with an integrator in the forward loop.

NAGD is a second-order algorithm with an additional pole $-\lambda$ and a zero $\frac{-\lambda}{1-\lambda}$.
A necessary condition for the stability of the algorithm is that all the poles of closed-loop system are within a unit disc.

If the Lipschitz continuous gradient constant $L$ is given, one can replace $A$ with $L$ and then the condition is sufficient.
For each $G(z)$, it has a corresponding iterative algorithm.

$G(z)$ can be third or higher order systems.

$G(z)$ can also be a fractional order system.
Analyses & Design with System Theory

General second-order algorithm design

Consider a general second-order discrete system

$$G(z) = \frac{z + b}{(z - 1)(z - a)}$$

whose corresponding iterative algorithm is

$$\begin{align*}
y_{k+1} &= ay_k - \mu \nabla f(x_k) \\
x_{k+1} &= x_k + y_{k+1} + by_k
\end{align*}$$

- Set $b = \frac{-a}{1+a}$ and one can derive the NAGD.
- Set $b=0$ and one can derive GDM.
The iterative algorithm can be viewed as the \textit{state-space realization} of the corresponding system. Thus, it has many different realizations (all are equivalent).

\[
\begin{align*}
Y(z) &= \frac{1}{z-a}[-\mu \nabla f(x_k)]_z \\
X(z) &= \frac{z-b}{z-1}Y(z) \\
Y(z) &= \frac{1}{z-1}[-\mu \nabla f(x_k)]_z \\
X(z) &= \frac{z-b}{z-a}Y(z)
\end{align*}
\sim
\begin{align*}
y_{k+1} &= ay_k - \mu \nabla f(x_k) \\
x_{k+1} &= x_k + y_{k+1} + by_k \\
y_{k+1} &= y_k - \mu \nabla f(x_k) \\
x_{k+1} &= ax_k + y_{k+1} + by_k
\end{align*}
\]
Analyses & Design with System Theory

General second-order algorithm design

We have introduced two parameters $a$ and $b$, but how to optimize them?

We can use *Integral Square Error (ISE)* as the criterion to optimize the parameters. Since for different target function $f(x)$, its Lipschitz continuous gradient constant is different. Thus, define $\rho := \mu A$ as the loop forward gain.
It is found that the optimal \( a \) and \( b \) satisfies
\[
b = \frac{-a}{1+a},
\]
which is the same design as NAGD.

We have used other criteria such as IAE, ITASE to find other optimal parameters, but the results are the same as ISE.

Different from NAGD, we derive the parameters by optimization rather than mathematically design, which can be extended to more general cases.
The MNIST database is a collection of hand-written digits, which contains 60,000 training images and 10,000 testing images. It is widely used as a benchmark for Machine Learning algorithms.

In the following, x-axis is always the epoch number.
Analyses & Design with System Theory

Figure: Training loss (Left), Test accuracy (Right)
Analyses & Design with System Theory

How does the zero influence the convergence performance?

- One can find that both $b = -0.25$ and $b = -0.5$ cases perform better than the SGD-Momentum. For $b = 0.25$ and $b = 0.5$, they perform worse.

- One can find the additional zero can improve the performance if we carefully adjust it.

- Both our method and Nesterov method give an optimal/good choice of the zero which is closely related to the pole ($b = \frac{-a}{1+a}$).
General third-order algorithm design

Consider a general second-order discrete system

$$G(z) = \frac{z^2 + cz + d}{(z - 1)(z^2 + az + b)}$$

- Set $b = d = 0$, it will reduce to the second-order algorithm.
- Compared with the second-order case, poles can now be complex numbers.
- More generally, a higher order system can contain **more internal models**.
Analyses & Design with System Theory

General third-order algorithm design

If all the poles are real, then one has that

\[ G(z) = \frac{1}{z-1} \left( \frac{z-c}{z-a} \right) \left( \frac{z-d}{z-b} \right) \]

whose corresponding iterative algorithm is

\[
\begin{align*}
y_{k+1} &= y_k - \mu \nabla f(x_k) \\
z_{k+1} &= az_k + y_{k+1} - cy_k \\
x_{k+1} &= bx_k + z_{k+1} - d z_k
\end{align*}
\]
## Analyses & Design with System Theory

### General third-order algorithm design (ISE)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.4</th>
<th>0.8</th>
<th>1.2</th>
<th>1.6</th>
<th>2.0</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.6439</td>
<td>0.5247</td>
<td>-0.4097</td>
<td>-0.5955</td>
<td>-1.0364</td>
<td>-1.4629</td>
</tr>
<tr>
<td>b</td>
<td>0.0263</td>
<td>0.0649</td>
<td>0.0419</td>
<td>-0.0398</td>
<td>0.0364</td>
<td>0.0880</td>
</tr>
<tr>
<td>c</td>
<td>1.5439</td>
<td>0.5747</td>
<td>-0.3763</td>
<td>-0.3705</td>
<td>-0.5364</td>
<td>-0.6462</td>
</tr>
<tr>
<td>d</td>
<td>0.0658</td>
<td>0.0812</td>
<td>0.0350</td>
<td>-0.0408</td>
<td>0.0182</td>
<td>0.0367</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.4</th>
<th>0.8</th>
<th>1.2</th>
<th>1.6</th>
<th>2.0</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roots of numerator</td>
<td>-1.5000</td>
<td>-0.3250</td>
<td>0.2082</td>
<td>0.1624</td>
<td>1.0000</td>
<td>1.4000</td>
</tr>
<tr>
<td></td>
<td>-0.0439</td>
<td>-0.2500</td>
<td>0.1681</td>
<td>0.6000</td>
<td>0.0364</td>
<td>0.0629</td>
</tr>
<tr>
<td>Roots of denominator</td>
<td>-0.6000</td>
<td>-0.3250</td>
<td>0.2128</td>
<td>0.1624</td>
<td>0.5000</td>
<td>0.5833</td>
</tr>
<tr>
<td></td>
<td>-0.0439</td>
<td>-0.1997</td>
<td>0.1969</td>
<td>0.3750</td>
<td>0.0364</td>
<td>0.0629</td>
</tr>
</tbody>
</table>
Analyses & Design with System Theory

General third-order algorithm design (ISE)

Since we use the ISE for tracking a step signal (it is quite simple), the optimal poles and zeros are the same as the second-order case with a pole-zero cancellation.

In this optimization results, all the poles and zeros are real.

Compared with the second-order case, the only difference is that we can have complex poles.

How to derive complex poles in the design? FC helps.
Analyses & Design with System Theory

General fractional-order algorithm design

Borrowing the idea from Micheal Jordan, we directly design a continuous time fractional order system

\[ G(s) = \frac{1}{s(s^\alpha + \beta)}, \alpha \in (0, 2), \beta \in (0, 20] \]

at first. Then, find the optimal parameters using the ISE criterion.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>1.8494</td>
<td>1.6899</td>
<td>1.5319</td>
<td>1.2284</td>
</tr>
<tr>
<td>( \beta )</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
Analyses & Design with System Theory

General fractional-order algorithm design

It is the continuous-time design. One can use the *Numerical Inverse Laplace Transform (NILT)* and Matlab command `stmcb()` to derive its discrete form.

\[
\begin{align*}
\rho &= 0.3 : G_3(z) &= \frac{0.0010z^2 + 0.0054z - 0.0032}{z^3 - 2.851z^2 + 2.767z - 0.9163}, \\
\rho &= 0.5 : G_3(z) &= \frac{0.0025z^2 + 0.0056z - 0.0042}{z^3 - 2.774z^2 + 2.625z - 0.8509}, \\
\rho &= 0.7 : G_3(z) &= \frac{0.0061z^2 + 0.0037z - 0.005}{z^3 - 2.652z^2 + 2.398z - 0.7462}, \\
\rho &= 0.9 : G_3(z) &= \frac{0.0187z^2 - 0.0082z - 0.0044}{z^3 - 2.357z^2 + 1.836z - 0.4788}, \\
\end{align*}
\]

\[
\begin{align*}
G_2(z) &= \frac{0.04689z - 0.04471}{z^2 - 1.958z + 0.9582}, \\
G_2(z) &= \frac{0.0426z - 0.03876}{z^2 - 1.927z + 0.9266}, \\
G_2(z) &= \frac{0.01962z - 0.00637}{z^2 - 1.748z + 0.7481}, \\
G_2(z) &= \frac{0.01895z - 0.0098}{z^2 - 1.434z + 0.4345},
\end{align*}
\]

where $G_2(z)$ is the second order approximation and $G_3(z)$ is the third order approximation.

https://www.mathworks.com/matlabcentral/fileexchange/39035-numerical-inverse-laplace-transform
If we direct design the algorithm in the discrete domain, all the poles are real.

The fractional order design contributes to the arise of complex poles.
Analyses & Design with System Theory

General fractional-order algorithm design

If we have complex poles, then one has that

\[
G(z) = \frac{z + c}{z - 1} \left( \frac{1}{z - a + jb} + \frac{1}{z - a - jb} \right)
\]

whose corresponding iterative algorithm is

\[
\begin{align*}
y_{k+1} &= ay_k - bz_k - \mu \nabla f(x_k) \\
z_{k+1} &= az_k + by_k \\
x_{k+1} &= x_k + y_{k+1} + cy_k
\end{align*}
\]
Analyses & Design with System Theory

General fractional-order algorithm design

We still use the MINST data set to compare the designed algorithm with the commonly used SGD algorithm and 50 epochs are used to train the Network. For the fractional order design, take the $\rho = 0.9$ case, where $a = 0.6786, b = 0.1354$, and different zero $c$ are designed. When $c = 0$, it is similar to the second-order SGD, while when $c \neq 0$, it is similar to second-order NAGD. For the SGD case, we set $\alpha = 0.9$. When simulation, set learning rate $\mu = 0.1$. 
Figure: $c = 0$: Training loss (Left), Test accuracy (Right)
Analyses & Design with System Theory

Figure: \( c = 0.283 \) Training loss (Left), Test accuracy (Right)
Analyses & Design with System Theory

- Both $c = 0$ and $c = 0.283$ cases perform better than commonly used SGD-Momentum.

- Generally, with $c$ carefully designed, better performance can be achieved as the second-order case.

- The simulation results can only prove that fractional calculus (complex poles) can potentially improve the performance, which is closely related to learning rate.
Observation and take home messages

Michael I. Jordan: *Is there an optimal way to optimize?* Yes, via limiting dynamics analysis and discretization, via SGD with other randomness like Langevin motion.

YangQuan Chen: *Is there a more optimal way to optimize?* Yes if Fractional Calculus is used:

- Optimal randomness in SGD, random search
- AGD limit dynamics is fractional order designed via FO E-L
- IMP
- … to be discovered
Outline

• Fractional Calculus, Complexity, and Fractional Order Thinking
• Big Data, Variability, and Fractional Calculus
• Machine Learning, Optimal Randomness and Fractional Calculus
• Looking Into Future: Fractional Calculus is Physics Informed
**Fractional Calculus: a response to more advanced characterization of our more complex world at too small or two large scale**

<table>
<thead>
<tr>
<th>1695</th>
<th>1960s</th>
<th>We are here</th>
</tr>
</thead>
<tbody>
<tr>
<td>static models</td>
<td>dynamical models</td>
<td>fractional order modeling</td>
</tr>
<tr>
<td>geometry, algebra</td>
<td>differential and integral calculus</td>
<td>fractional calculus</td>
</tr>
</tbody>
</table>

Slide credit: Igor Podlubny
Take home messages

Want to do **better than the best?**
Want to be **more optimal?**

Go Fractional!
Decision and Control in the Era of Big Data?

• Yes, we must use fractional calculus!
  – Fractional order signals, systems, controls.
  – **Fractional order data analytics**
Future of Machine Learning

- Physics-informed ML
- Scientific ML – (cause-effect embedded or cause-effect discovery)

- Involving fractional calculus, we are closer to the nature, i.e., “道”
New wisdom equipped with FC

- 道可道，非常道。世间万物的运行规律是可以被描述的，但它们并非一成不变的。

Non-normal way: Fractional Calculus! Heavytailedness

Why BD and ML must meet FC?
New wisdom equipped with FC

- 玄之又玄，众妙之门。了解这类对立统一相互转变的规律，就是通向对世间万物理解的大门。

Root of long (algebraic) tail, or inverse power law

\[
E_{\alpha,1}(-x^\alpha) \approx \begin{cases} 
1 - x^\alpha / \Gamma(1+\alpha) & x \to 0^+ \\
 x^{-\alpha} / \Gamma(1-\alpha) & x \to \infty 
\end{cases}
\]

Non-normal way:

**Fractional Calculus! Heavytailedness**

Why BD and ML must meet FC?
To have a better life, learn FC

• 老子说：“人法地，地法天，天法道，道法自然。”
  – “道法自然” – prompts the use fractional calculus

• Better understanding complexity using fractional calculus leads to 积极入世的态度（王阳明）then “天人合一”
  – 天人不合一例子：逆水行舟，冒雨走夜路，冬天穿背心。。。 [1]
  – 天人合一例子：夫妻恩爱，团队精神，人养狗、狗护主。。。

07/06/2020  Why BD and ML must meet FC?
Q/A session

Thank you for your attendance and patience! Your comments and critiques are welcome!

• Thanks go to
  – Dr. Yuquan Chen, Dr. Jiamin Wei, Lihong Guo, Dr. Zhenlong Wu, Dr. Yanan Wang, Panpan Gu, Jairo Viola, Haoyu Niu, Dr. Jie Yuan etc. for walks, chats and tea/coffee breaks at Castle, Atwater, CA before COVID-19 era.
Backup slides
• Socrates,
  – https://en.wikipedia.org/wiki/The_unexamined_life_is_not_worth_living
• Platos, Aristotle

Pythagoras  
C. 570 – c. 495 BC

Parmenides  
Around 475 BC

Heraclitus  
C. 535 – C. 475 BC

07/06/2020

Why BD and ML must meet FC?
Fractional Order Controls

• IO Controller + IO Plant
• FO Controller + IO Plant
• FO Controller + FO Plant
• IO Controller + FO Plant


Why BD and ML must meet FC?