THE OTHER SIDE OF APPLIED MATHEMATICS: SOME DIFFUSE REFLECTIONS

Vivek Borkar

IIT Bombay

Jan. 23, 2017, ISI Bengaluru

The world according to Mark

- World War II $\Longrightarrow$ rise in prominence of aerodynamics, fluid mechanics, solid mechanics, $\cdots$

- These and allied areas promoted in British science (Lighthill), come to dominate applied mathematics in UK

- Relatively less emphasis on other aspects of applied mathematics, leading to their decline

(Fallout in India.)

Good reasons for the rise of numerical ODE and PDE, fluid mechanics, $\cdots$:

Classical ‘core’ engineering areas: mechanical, civil,

chemical, electrical power, metallurgy, aerospace

Deal with: materials and energy generation, storage, distribution, utilization

‘Physics’ of the tasks calls for the mathematics of above flavor

$\Longrightarrow$ these strands of mathematics form the core ‘engineering mathematics’ courses

(cf. books by Thomas, Kreyszig)

Typical fare: matrices, multivariable calculus, basic complex analysis, transforms, linear ODE solved by transform techniques or by plugging in series, elementary PDE (e.g., separation of variables), some discrete probability

The other side of engineering

Generation, storage, distribution and utilization

of information $\approx$ abstract symbols (‘signals’)

Primary customers:

Electrical engineering: control and communications

Computer science and information technology

Industrial engineering and operations research

ISSUES:

1. Training for these is available in courses that are at best scattered and not cohesively organized. It is rare that such tracks are even flagged as a legitimate strand.

2. Not much guidance or direction is available for interested students at undergraduate level, exposure is often left to chance. Paucity of expository or popular material.

Consequence:

1. low awareness of modern developments in many of these areas among students and faculty, major developments passing us by

2. activity mainly in engineering departments with more utilitarian focus. In comparison, the activity in mathematics community is low in numbers, often ‘fossilized’ $\approx$ trapped in a time bubble, expended on contrived, tangential problems **(More on this later)**

3. high quality activity on these themes in mathematics community is sporadic and scattered, hardly a community. Falls between two stools: not much support from either the pure mathematics community or the traditional applied mathematics community

Three examples:

1. Crawling for ephemeral content

2. ‘Gossip’ type algorithms

3. Rumour source detection

Problem 1:
Crawling for ephemeral content

(joint with K. Avrachenkov, INRIA,
supported by IFCPAR)

Some news items:

- 'Just for animals' terminal at JFK Airport
- Half of Britain does not believe in God
- Messi motivates me to scale greater heights: Ronaldo
- Cyrus Broacha bonds with son over cricket

‘Ephemeral content’:

Web content of immediate interest, but interest rapidly decays with time.

Problem:

How to schedule web crawlers to capture ephemeral web content?

THE MODEL (Empirically validated)

1. $X_i(n) :=$ ‘web content’ at location i at time n, $1 \leq i \leq N$.

2. $\alpha_i \in (0, 1)$ decay rate of ‘interest’

3. $u_i :=$ mean arrival rate of ‘content’ per epoch

Then the dynamics is:

$X_i(n+1) = \alpha_i X_i(n) + u_i$ if not crawled,

$X_i(n+1) = u_i$ if crawled.

Control variable: $v_i(t) = 1$ if crawled, 0 otherwise.

Objective: Maximize average reward

$$\limsup_{t\uparrow\infty} \sum_{i=1}^{N} \frac{1}{t} \sum_{m=0}^{t} X_i(m) v_i(m)$$

subject to

$$\sum_{i=1}^{N} v_i(m) = M \quad \forall\ m \geq 0.$$

This problem is hard, so use the Whittle relaxation:

(per stage constraint → time-averaged constraint)

$$\lim_{t \uparrow \infty} \sum_{i=1}^{N} \frac{1}{t} \sum_{m=0}^{t} v_i(m) = M.$$

This is an instance of '*restless bandits*'.

Multi-armed bandits:

N processes (Markov chains), out of which $M < N$ can be operated at a time ('active' arms), while the rest remain *frozen* ('passive' arms).

Problem:

Optimal scheduling

Typical solution: ‘index rule’ (Gittins): to each process is assigned a state-dependent index, use the bandit with the maximum index. (Optimal)

Restless bandits: Passive bandits drift according to a neutral dynamics.

Only a heuristic is available, even after relaxing the rigid condition of ‘M out of N’ to ‘M out of N on average’.

Whittle index:

Consider the 'M out of N on average' version.

For each process i, introduce subsidy λ_i for remaining passive.

If the set of states at which it is optimal to remain passive monotonically increases from 'none' to 'all' as the subsidy increases from $-\infty$ to ∞ for each i, the problem is *Whittle indexable*.

Whittle index of $i =$ the λ_i at which active and passive are equally desirable, as a function of state.

Index policy: Operate top M according to diminishing order of indices

This is suboptimal, but asymptotically optimal as $N \uparrow \infty$ (Weiss-Weber).

Known to perform well in practice.

Some applications:

Sensor scheduling

Multi-UAV coordination

Congestion control

Cognitive radio

Real time wireless multicast

Intuition:

- Consider the 'M out of N on the average' problem with the latter cast as an additional average cost constraint.
- This makes it a 'constrained Markov decision process'.
- Important special feature: separable cost and separable constraint.

- Consider the standard ‘LP formulation’ in terms of occupation measures: Maximize

$$\sum_i \int f_i d\mu_i \text{ s.t. } \sum_i \int g_i d\mu_i \leq C.$$

- Use Lagrange multiplier to formulate the equivalent unconstrained problem: Maximize

$$\sum_i \int (f_i - \lambda g_i) d\mu_i.$$

- Separability of cost and constraint $\Longrightarrow$ given the Lagrange multiplier λ, the problem decouples

- Each separate average cost control problem involves a binary decision variable:

  to be or not to be (active / passive)

  $\Longrightarrow$ decision boundary comes from an inequality of the type 'a function of state $\leq \lambda$' specifying the passive states.

- Whittle indexability $\Longrightarrow$ the feasible set of this inequality (passive set) increases from empty set to the whole space as $\lambda \uparrow \infty$.

- This suggests that for any single process, the value of λ for which active and passive modes become equally desirable is a measure of selection for activity.

(λ can be interpreted as ‘subsidy for passivity’)

- Set ‘Whittle index’ := the value of λ for which this equality is achieved, as a function of state for each process.

- Whittle’s heuristic: Choose the top M according to decreasing value of indices for the current profile of state variables.

Back to crawling

For the crawling problem, let

$$\zeta_i(x) := \left\lceil \log^{+}_{\alpha_i}\left(\frac{u_i-(1-\alpha_i)x}{\alpha_i u_i}\right)\right\rceil .$$

Then the Whittle index is (for non-boundary cases)

$$\gamma_i(x) := (1{+}\zeta_i(x)((1{-}\alpha_i)x{-}u_i){+}\left[\alpha_i^{\zeta_i(x)} + \left(\frac{1-\alpha_i^{\zeta_i(x)}}{1-\alpha_i}\right)\right] u_i .$$

'Boundary cases' (always/never crawl) can be treated separately.

Proof technique:

1. Consider a separate unconstrained control problem with subsidy for each i.

2. Consider ‘discounted reward’ $\sum_n \beta^n E[\cdots]$, $\beta \in (0,1)$, and establish the corresponding dynamic programming equation.

3. Justify the ‘average reward dynamic programming equation’ for each i by the ‘vanishing discount’ argument. (Uses ‘coupling’ argument.)

4. Check that the corresponding ‘value function’ is monotone increasing and convex, and also has the ‘increasing differences’ property:

$$V(\lambda + a, x + b) - V(\lambda + a, x) \geq V(\lambda, x + b) - V(\lambda, x)$$

for $a, b > 0$.

5. Get *threshold* policy: passive up to a level, then active, with threshold monotone with λ.

$\Longrightarrow$ Whittle indexability.

6. Use the definition of Whittle index and the dynamic programming equation to derive the Whittle index as a function of state.

   (Not always possible $\Longrightarrow$ need computational schemes for its approximate evaluation of Whittle indices.)

Problem 2:
‘Nonlinear’ gossip

(joint with A. Mathkar, Goldman Sachs,
supported by DST)

STOCHASTIC APPROXIMATION

Consider the Robbins-Monro scheme in $\mathcal{R}^d$:

$$x(n+1) = x(n) + a(n)[h(x(n)) + M(n+1)].$$

Here:

- $h : \mathcal{R}^d \mapsto \mathcal{R}^d$ Lipschitz,

- $\{M(n)\}$ a martingale difference sequence w.r.t. $\mathcal{F}_n := \sigma\left(x(m), M(m), m \leq n\right), n \geq 0$, i.e.,

$$E\left[M(n+1)|\mathcal{F}_n\right] = 0.$$

Also, there exists $K \in (0, \infty)$ such that

$$E\left[\|M(n+1)\|^2 | \mathcal{F}_n\right] \leq K\left(1 + \|x(n)\|^2\right).$$

- Step-sizes $a(n) > 0$ satisfy:

$$\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty.$$

‘ODE Approach’ (Derevitskii-Fradkov-Ljung)

View the iteration as a noisy discretization of the ODE

$$\dot{x}(t) = h(x(t)), \; t \geq 0.$$

This is well posed under our hypotheses.

Definition: A set A is invariant if

$$x(0) \in A \Longrightarrow x(t) \in A \; \forall \; t \in \mathcal{R}.$$

Definition (continued):

A is *Internally Chain Transitive* if given any $x, y \in A$, and $\epsilon > 0, T > 0$, we can find $n \geq 1$, and

$$x_0,\ x_1,\ \cdots,\ x_{n-1},\ x_n = y \in A$$

such that $\|x - x_0\| < \epsilon$ and for $0 \leq i < n$, the trajectory $x^i(t), t \geq 0$, of

$$\dot{x}^i(t) = h(x^i(t)),\ x^i(0) = x_i,$$

satisfies $\|x^i(t) - x^{i+1}\| < \epsilon$ for some $t \geq T$.

Benaim's theorem:

If $\sup_n \|x(n)\| < \infty$ a.s., then $x(n) \to$ a compact connected nonempty internally chain transitive invariant set of the ODE, a.s.

(Starting point for finer results using problem specifics)

THE TSITSIKLIS MODEL

- ‘Agents’/processors placed at the nodes of an irreducible directed graph $\mathcal{G}$ with node set $\mathcal{V}$ with $|\mathcal{V}| := N$ and edge set $\mathcal{E}$. $\mathcal{N}(i) := \{i\text{'s neighbors}\}$.

- For $i \in \mathcal{V}$ and $P = [[p(j|i)]]$ stochastic, $\mathcal{G}$-compatible,

$$x_i(n+1) = \sum_j p(j|i)x_j(n) + a(n)[h(x_i(n)) + M_i(n+1)].$$

- At each instant, every node takes,

  – a weighted average of its neigbhbors' values (**'gossip' component**), and,

  – adds a correction based on its own computation (**'learning' component**).

- Delays, asynchrony, etc. (shall worry about it later).

Similar models (albeit in continuous time, possibly 'second order') in synchronization, flocking/coordination, $\cdots$.

Objective: **CONSENSUS**

Usual approaches:

- Product of (possibly random) stochastic matrices (Chatterjee-Seneta, etc.)
- ‘Coupling from the past’

Alternative viewpoint:

- View the iteration as a slowly vanishing regular perturbation of vanilla gossip $\Longrightarrow$ two time scales

- ‘Gossip’: a marginally stable system with one dimensional invariant subspace (= the Perron–Frobenius eigenvector)

- Convergence to the invariant subspace (consensus) + selection via the slower dynamics

I: quasi-linear case

For each $i \in \mathcal{V}$, consider the d-dimensional iteration

$$\begin{aligned} x_i(n+1) \;=\; & \sum_{j \in \mathcal{N}(i)} p_{x(n)}(j|i) x_j(n) + \\ & a(n)\left[h_i(x_i(n)) + M_i(n+1)\right]. \end{aligned}$$

Here, P_x is an irreducible stochastic matrix where $x \mapsto P_x$ is Lipschitz, with $(\min)_j^+ p_x(j|i) \geq \Delta > 0$.

For a fully distributed algorithm, the ith row of $P_{x(n)}$ should depend only on $x_j(n),\ j \in \mathcal{N}(i) \cup \{i\}$.

We use $x(n)$ without loss of generality.

Can also have $h_i(x(n))$ instead of $h_i(x_i(n))$.

Let $\pi_x :=$ the unique stationary distribution under P_x.

CONSENSUS:

if $\sup_{i,n} \|x_i(n)\| < \infty$ a.s., then

$$\|x_i(n) - x_j(n)\| \to 0 \text{ a.s.}$$

(Not surprising, standard arguments work.)

MAIN RESULT ($d = 1$):

Let $\mathcal{A} := \{c\mathbf{1} : c \in \mathcal{R}\}$. Let $x(n) = [x_1(n), \cdots, x_N(n)]^T$.

If $\sup_{i,n} \|x_i(n)\| < \infty$ a.s., then almost surely, $x(n) \to \mathcal{A}_0 :=$ an internally chain transitive invariant set of N-fold copy of the ODE

$$\dot{y}(t) = \sum_k \pi_{y(t)\mathbf{1}}(k) h_k(y(t)), \ t \geq 0,$$

contained in $\mathcal{A}$.

General case: Define

$$\mathcal{A} \ := \ \{x = [(x^1)^T : \cdots : (x^N)^T]^T \in \mathcal{R}^{d\times N} : \\ x^i = [x^i_1, \cdots, x^i_d]^T, 1 \le i \le N; \ x^i_k = x^j_k \ \forall \ i, j\}.$$

Consider

$$\dot{y}(t) = \sum_{i=0}^{N} \pi_{\psi(y(t))}(i) h_i(y(t)).$$

where $\psi(y) := [y^T : y^T : \cdots : y^T]^T$ for $y \in \mathcal{R}^d$.

Then $\mathcal{A}$ is invariant under this dynamics.

Theorem $\text{sup}_n \|x_n\| < \infty$ a.s. $\Longrightarrow x(n) \overset{n\uparrow\infty}{\to}$ a compact connected non-empty internally chain transitive invariant set $\mathcal{A}_0 \subset \mathcal{A}$ of the N-fold product of the above dynamics, a.s.

(That is, dynamics in $\mathcal{R}^N$ wherein each component satisfies the above o.d.e.)

Stronger results possible for special cases
(e.g., convergence for $d = 1$!)

Example: Consider $h_i = -\nabla f \ \forall i$. Let $|\mathcal{N}(i)| = M \ \forall i$ and for a prescribed $T > 0$ ('temperature')

$$
\begin{aligned}
p_x(j|i) &= \frac{1}{M} e^{-\frac{(f(x_j)-f(x_i))^+}{T}}, \ j \in \mathcal{N}(i), \\
&= 0, \qquad\qquad j \notin \mathcal{N}(i), j \neq i, \\
&= 1 - \sum_{k \in \mathcal{N}(i)} p_x(k|i), \quad j = i.
\end{aligned}
$$

Then

$$\pi_x = \frac{e^{-\frac{f(x_i)}{T}}}{\sum_j e^{-\frac{f(x_j)}{T}}}.$$

This puts more weight on low values of f (spatial annealing).

Can think of this scheme as a '*leaderless swarm*' by analogy with *Particle Swarm Optimization*, wherein each particle uses information from self, neighbors, and the 'best so far', i.e., a leader. Here the last piece is 'emergent' from a distributed gossip.

Another example: Dependence of P_x on x due to mobility.

Some intuition:

Think of this as a two time scale phenomenon:

∘ gossip on fast ‘natural’ time scale $n = 0, 1, 2, \cdots\cdots$, and,

∘ learning on the slow ‘ODE’ time scale:
$t(0) = 0, t(1) = a(0), \cdots, t(n) = \sum_{i=0}^{n-1} a(i), \cdots$.

Compare with traditional two time scale schemes:

$$\begin{aligned} x(n+1) &= x(n) + a(n)[h(x(n), y(n)) + M(n+1)], \\ y(n+1) &= y(n) + b(n)[g(x(n), y(n)) + M'(n+1)], \end{aligned}$$

with $b(n) = o(a(n))$.

In contrast, the two time scales are now a part of the same iteration.

(akin to ‘Markov noise’:

$x(n+1) = x(n) + a(n)[h(x(n), Y(n)) + M(n+1)]$.)

Also, a *‘stability test’*: Define

$$\begin{aligned} g(x) &:= \sum_i \pi_x(i) h_i(x), \\ g_c(x) &:= \frac{g(cx)}{c} \text{ for } c > 0, \\ g_\infty(x) &:= \lim_{c \uparrow \infty} g_c(x), \end{aligned}$$

assumed to exist. Then g_c, g_∞ are Lipschitz.

Consider the ODE (‘scaling limit’)

$$\dot{x}_\infty(t) = g_\infty(x_\infty(t)), \ t \geq 0.$$

If this has the origin as the unique asymptotically stable equilibrium, then $\sup_n \|x(n)\| < \infty$ a.s.

Intuition: Iterates large in absolute value track this o.d.e. after scaling, hence exhibit stabilizing drift.

II: fully nonlinear case

For each $i \in \mathcal{V}$, consider the d-dimensional iteration

$$x_i(n+1) = f_i(x(n)) + a(n)\left[h_i(x_i(n)) + M_i(n+1)\right].$$

Here:

- $f := [f_1, \cdots, f_N]^T : (\mathcal{R}^d)^N \mapsto (\mathcal{R}^d)^N$ is continuous, and,

- $P(x) = \lim_{n\uparrow\infty} f^{(n)}(x)$ ($:= f \circ f \circ \cdots \circ f$, n times) exists, with the limit being uniform on compacts.

  Then

$$\begin{aligned} P(P(x)) &= P(f(x)) \\ &= f(P(x)) \\ &= P(x) \\ &\in C := \{x : P(x) = x\}. \end{aligned}$$

Assumptions:

1. P is Frechet differentiable with its Frechet derivative $\bar{P}_x(\cdot)$ continuous in x.

2. $\bar{P}_{f(\cdot)}h(\cdot)$ is Lipschitz. (Ideally, should be ‘local’, but we ignore this issue.)

3. $E\left[\|M(n+1)\|^4|\mathcal{F}_n\right] \leq F(x(n))$ for some continuous F.

Assume $\sup_n \|x(n)\| < \infty$ a.s.

Consider the ODE

$$\dot{x}(t) = \bar{P}_{x(t)}(h(x(t))).$$

MAIN RESULT: $x(n) \rightarrow$ a compact connected nonempty internally chain transitive invariant set of the above ODE contained in C, a.s.

Example: $P :=$ a projection to a convex set, $x(n+1) = f(x(n))$ an iterative scheme for calculating the projection.

In this case, we get a projected version of the distributed stochastic approximation scheme.

$\Longrightarrow$ Need distributed scheme for computing projections on, e.g., intersection of convex sets.

(More on this later.)

Standard issues in distributed computation:

1. Interprocessor delays

2. Asynchrony: not all updates at the same time

3. Updates may be on ‘local clock’

Under suitable modifications, earlier results hold:

1. Bounded delays ‘squeezed out’ (i.e., they lead to asymptotically negligible error) due to time scaling (more generally, conditional moment conditions suffice)

2. Asynchrony / local clocks compensated for by the choice of stepsize (get back the original limiting ODE modulo time-scaling)

Application: Projected stochastic approximation in a convex set C given as intersection of convex sets $\{C_i\}$

Classical approach: projection at each step, leads to the limiting ODE

$$\dot{x}(t) = \bar{\Pi}_{x(t)}(h(x(t))).$$

where $\bar{\Pi}$: Frechet derivative of the projection operator (differential inclusion in case of non-smooth boundaries) (Kushner-Clark)

Objective: distributed scheme where node i has access only to C_i

- ‘Fast’ time scale: iterative scheme for computing the projection to C in a distributed manner

  (Distributed Boyle-Dykstra-Han algorithm,
  joint work with Soham Phade)

- ‘Slow’ scale: stochastic approximation

Combined scheme can be shown to be stable and convergent with probability one, and tracks the projected stochastic approximation scheme as desired
(joint work with Suhail Shah)

Problem 3:

Rumour source detection

(joint work with Ankit Kumar, Samsung,

and Nikhil Karamchandani, IITB,

supported by IFCPAR)

(Shah and Zaman '11, '12)

Problem: find the source of rumour given a snapshot of how far it has spread.

Shah-Zaman approach:

- SI model of spread.
- exact MLE for tree graphs (‘rumour centrality’).

- Apply the same to BFS trees in general graphs.

- Effective heuristics for sparse graphs.

- Sophisticated asymptotics for random graphs.

Alternative model:

- Partial pairwise influence data available.
- Assume all compatible spread patterns equally likely.
- Find the node with the largest number of compatible spread patterns.

Our approach:

- Use **Markov chain tree theorem**: stationary distribution of a node is proportional to sum of weights of arborescences* rooted in it, where the latter is the product of edge weights (transition probabilities) associated with the arborescence

*maximal rooted trees with at most one outgoing edge per node

- Stationary distribution for random walk on a connected undirected graph is proportional to its degree

  $\Longrightarrow$

- Ranking based a quantity proportional to the number of arborescences rooted at a node $\left(\propto \frac{\pi(i)}{d(i)} \right)$.

  Nontrivial pre-processing (pruning) required.

- adapt Aldous's algorithm based on Markov Chain Monte Carlo (MCMC) for counting spanning trees in order to estimate the rumour source

(Another option: Wilson's algorithm)

Good performance observed for :

- dense graphs (Shah-Zaman scheme based on ‘most likely (BFS) tree’, works well in the sparse case)

- mismatched models (e.g., ‘Big Fat Tree’ algorithm for IC model applied to SI model)

Reason: model-free, uses only counting (does not use timing information)

Extension to the harder case when information non-local, i.e., of the form 'i infected before j' where i, j may not be neighbours: based on rejection sampling.

Above scheme slow, currently trying an 'arborescence-valued MCMC' based on a construct of Anantharam and Tsoucas, with matrix analytic gimmicks such as the Sherman-Morrrison-Woodbury formula to simplify computations. Gives promising results.
(joint with Anand Kalvit, Nikhil Karamchandani)

CODA:

On being trapped in a time bubble –

- There's more to computational mathematics than plotting the flow between a cylinder with octahedral cross-section rotating inside another with oblong cross-section

  (computational topology, high dimensional optimization, ...)

- There's more to probability than numerical approximation of the transient response of two queues in tandem.

  (queuing networks, limiting measure-valued processes, …)

- There's more to optimization than trivial extensions of KKT conditions to, say, 'asymptotically lower semi-(α, β, γ)-quasi-convex' functions[†].

  (semi-definite programming, polynomial optimization, ...)

- There's more to theory of computation than fuzzy compact operators applied to the model of a traffic junction.

  (kernel methods, reinforcement learning, ...)

[†] not to be confused with 'supercalifragilisticexpialidocious'

- There's more to graph theory than properties of vertex-crazy edge-indifferent hypergraphs.

  (random graphs, graph limits, ...)

Though exaggerated, these 'examples' are not too far off the mark, as evidenced by submissions one gets for our mathematics journals.

A PARTING MESSAGE[‡]:

- new collection of ideas/problems emerging because of changing technological landscape, e.g., around the broad theme of ‘networks’ (internet, world wide web, social networks, sensor networks, robotic swarms, ‘big data’, $\cdots$)

- draws upon unconventional mixes of methods from linear algebra, (high dimensional) probability, statistics, optimization, algorithmic complexity, $\cdots$

[‡] cf. Nevanlinna Award speech of Jon Kleinberg

- deserves to be flagged as a distinct new domain of applied mathematics with associated coursework and programs

Unless we nurture these under-represented strands of applied mathematics, several important modern developments will pass us by.

And while we were busy admiring the spring blossoms,
the caravan passed us by and we were left staring at
the dust in its wake.

– Neeraj (freely translated from Hindi)

THANK YOU

and Best Wishes to Ram !