doi:10.20944/preprints202012.0712.v1

Forecasting COVID-19 with importance-sampling and path-integrals

Lester Ingber
Physical Studies Institute LLC, Ashland, OR, USA
ingber@caa.caltech.edu

Abstract

Background: Forecasting nonlinear stochastic systems most often is quite difficult, without giving in to temptations to simply simplify models for the sake of permitting simple computations.
Objective: Here, two basic algorithms, Adaptive Simulated Annealing (ASA) and path-integral codes PATHINT/PATHTREE (and their quantum generalizations qPATHINT/qPATHTREE) are suggested as being useful to fit COVID-19 data and to help predict spread or control of this pandemic. Multiple variables are considered, e.g., potentially including ethnicity, population density, obesity, deprivation, pollution, race, environmental temperature.
Method: ASA and PATHINT/PATHTREE have been demonstrated as being effective to forecast properties in three disparate disciplines in neuroscience, financial markets, and combat analysis.
Results: Not only can selected systems in these three disciplines be aptly modeled, but results of detailed calculations have led to new results and insights not previously obtained.
Conclusion: While optimization and path-integral algorithms are now quite well-known (at least to many scientists), these applications give strong support to a quite generic application of these tools to stochastic nonlinear systems.

Keywords: path integral, importance sampling, financial options, combat analysis

1. Introduction

It is generally recognized that the spread of COVID-19 is affected by multiple variables, e.g., potentially including ethnicity, population density, obesity, deprivation, pollution, race, environmental temperature (Anastassopoulou *et al*, 2020; Bray *et al*, 2020; Li *et al*, 2020). Also, the Centre for Evidence-Based Medicine (CEBM) regularly cites papers on the dynamics of COVID-19 at https://www.cebm.net/evidence-synthesis/transmission-dynamics-of-covid-19/ .

This proposal offers the application of two basic multivariate algorithms to fairly generic issues in forecasting. As such, they may be useful to fit COVID-19 data and to help predict upcoming spread and control of this pandemic.

> (a) Adaptive Simulated Annealing (ASA) is an importance-sampling optimization code usually used for nonlinear, nonequilibrium, non-stationary, multivariate systems.
>
> (b) PATHINT is a numerical path-integral PATHINT code used for propagation of nonlinear probability distributions, including discontinuities.

These codes were developed by the author and applied across multiple disciplines.

There is not "one size fits all" in forecasting different systems. This was demonstrated for three systems (Ingber, 2020b), where the author has addressed multiple projects across multiple disciplines using these tools: 72 papers/reports/lectures in neuroscience, e.g. (Ingber, 2018; Ingber, 2020c), 31 papers/reports/lectures in finance, e.g. (Ingber & Mondescu, 2003; Ingber, 2020a), 24 papers/reports/lectures in combat analyses, e.g. (Ingber, 1993; Ingber, 2015), and 11 papers/reports/lectures in optimization, e.g. (Atiya *et al*, 2003; Ingber, 2012), It is reasonable to expect that this approach can be applied to many other projects.

For example, the path-integral representation of multivariate nonlinear stochastic differential equations permits derivation of canonical momenta indicators (CMI) which are faithful to intuitive concepts like Force, Momenta, Mass, etc (Ingber, 1996; Ingber, 2015; Ingber & Mondescu, 2001). Correlations among variables are explicitly included in the CMI.

2. Data

A large and updated database for COVID-19 is maintained by the John Hopkins University (JHU) at https://github.com/CSSEGISandData/COVID-19/blob/master/archived_data/archived_daily_case_updates/01-21-2020_2200.csv . This database was used for a pilot study.

2.1. 50+ Locations

The data being used contains 3340 cities throughout the US and some territories. The locations have been broken into 57 States and Territories ready for production runs.

3. Technical considerations

If there is not time to process large data sets, then the data can be randomly sampled, e.g., as described in another paper, “Developing bid-ask probabilities for high-frequency trading” (Ingber, 2020a).

If the required forecast is longer than the conditional distribution can sustain, PATHINT/PATHTREE can be used to propagate the distribution.

The dataset should be broken into independent Training and Testing subsets, to test the trained distribution. If this is not possible, e.g., because of data or time limitations, at the least experts can be used to judge if the model is ready for real-time applications, e.g., the Delphi method (Okoli & Pawlowski, 2004).

If an algorithm like ASA is to be used across a large class of problems, then it must be tunable to different classes. Over the 30+ years of ASA development, the author has worked with many volunteers who have contributed valuable ideas, modifications and corrections to this code. This has resulted in over 150 ASA options that can be used for additional timing additional tuning making it useful across many classes of problems.

The path integral algorithm includes its mathematical equivalents, a large class of stochastic differential equations and a large class of partial differential equations. The advantages of the path integral algorithm are:

> (a) intuitive description in terms of classical forces, inertia, momentum, etc., leading to new indicators.
>
> (b) delivering a cost function derived from a Lagrangian, or its Action (Lagrangian x dt). Sometimes constraints need to be added as Lagrange multipliers, as was required for normalization requirements in financial risk projects (Ingber, 2010).

4. Pilot Study

The shape of the spread of this virus is clearly nonlinear. A simple model was used for a pilot study to at least capture some nonlinearity. For example, just using the daily number of total cases reported, C, the short-time conditional Probability $P(t+1|t)$ is given in terms of its effective Lagrangian L, $P = \exp(-L_{eff}\,dt)$ (including the logarithm of the prefactor normalization as it may contain nonlinearities as modeled here):

$$
\begin{aligned}
L_{eff} &= [(x_{t+1} - x_t - g_x dt) g_{xx'} (x'_{t+1} - x'_t - g_{x'} dt) + 1/2 \log(2\pi dt g^2) \\
g_x &= a \exp(x^b) \\
g_{xx'} &= c \exp(x^d) \\
g &= \det(g_{xx'}) \qquad (1)
\end{aligned}
$$

with parameters to be fit to data $\{a, b, c, d\}$. This is a simple one-factor model. In more than one dimension, $g_{xx'}$ is the metric of this space, the inverse of the covariance matrix.

1000 acceptance iterations of this cost/objective function, taking 6247 generated states, over the JHU data gave

$$a = 0.077 , b = 0.874 , c = 2.79 , d = 0.845 \qquad (2)$$

4.1. Comet Profile

These codes were run on XSEDE Comet, for just 1000 generated states, to profile the current code.

"Comet is a dedicated XSEDE cluster designed by Dell and SDSC delivering 2.0 petaflops, featuring Intel next-gen processors with AVX2, Mellanox FDR InfiniBand interconnects and Aeon storage. The standard compute nodes consist of Intel Xeon E5-2680v3 (formerly codenamed Haswell) processors, 128 GB DDR4 DRAM (64 GB per socket), and 320 GB of SSD local scratch memory. The GPU nodes contain four NVIDIA GPUs each. The large memory nodes contain 1.5 TB of DRAM and four Haswell processors each. The network topology is 56 Gbps FDR InfiniBand with rack-level full bisection bandwidth and 4:1 oversubscription cross-rack bandwidth. Comet has 7 petabytes of 200 GB/second performance storage and 6 petabytes of 100 GB/second durable storage. It also has dedicated gateway hosting nodes and a Virtual Machine repository. External connectivity to Internet2 and ESNet is 100 Gbps."

Comet is being phased out and users will soon be using the new Expanse platform.

4.2. Timing Runs

1000 acceptance iterations of this cost/objective function, taking 6247 generated states, over the JHU data gave

$$a = 0.077 , b = 0.874 , c = 2.79 , d = 0.845$$

The time to process the data on the author's Thinkpad P1 Gen 3 with a Xeon CPU running Windows Pro Workstation x64, using 'gcc -O3', determined using '/bin/time -p make' was:

```
real 256.73
user 254.85
sys 0.79
```

On Comet, also using 'gcc -O3', gave:

```
real 377.42
user 377.32
sys 0.05
```

4.3. Future Runs

An interesting aspect of the above simple model is that powers of exponential behavior of evolving cases of the virus can be simply calculated. This may give some insight into the spread rates, e.g., the effective number of contacts made by people with this disease.

Note that production runs likely will include more complex models and longer runs, requiring more resources than this pilot study.

4.4. Parallel Processing

"Parallel Processing for this project basically is similar to many projects developed by the author as Principal Investgator at the Extreme Science and Engineering Discovery Environment (XSEDE.org) since February 2013. That is "trivial MPI" is used, wherein many simulataneous runs are achieved by simply reading in different data files to ASA, using the "array" feature offered by some XSEDE platforms. As offered in a previous XSEDE Extended Collaborative Support Service (ECSS) ticket:

> Parallelization efficiency is 1 for jobs running on a single core that is max one could get. For multi-threaded apps one can get some to decent bump in speed using multiple cores up to some point before plateauing. However, speed bump with multiple cores often leads drop in parallelization efficiency.
>
> Drawback of using single core is too long run time. Though in this case, you are running array jobs with single core and getting maximum efficiency. This is the ideal situation on

doi:10.20944/preprints202012.0712.v1

> ’Comet’ because nodes on this machine can be shared. You should explain on Scaling and parallelization efficiency section that your application is not multi-threaded and you use single core on comet to run your jobs. This gives efficiency of 1, which is maximum value achievable. However, you run array of jobs in one submission and each job uses a single core. This is most efficient use of resources because node sharing is allowed on Comet. It won’t hurt to write that you have consulted XSEDE staff on this matter.”

5. Conclusion

Two algorithms are suggested for fitting data and forecasting COVID-19, ASA for importance-sampling and fitting parameters to models, and PATHINT/PATHTREE. These algorithms have been applied to several disciplines — neuroscience, financial markets, combat analysis. While optimization and path-integral algorithms are now quite well-known (at least to many scientists), these previous applications give strong support to application of these tools to COVID-19 data.

Acknowledgments

The author thanks the Extreme Science and Engineering Discovery Environment (XSEDE.org), for supercomputer grants since February 2013. The PI has held yearly grants for supercomputer resources from XSEDE, mostly for work on SMNI but for a couple years also to develop qPATHINT and qPATHTREE codes for propagation of quantum wave functions in the presence of random shocks.

References

Note that some URLs are cited in-text.

C. Anastassopoulou, L. Russo, A. Tsakris & C. Siettos (2020) Data-based analysis, modelling and forecasting of the COVID-19 outbreak. Public Library of Science One. 15(3). [URL https://doi.org/10.1371/journal.pone.0230405]

A.F. Atiya, A.G. Parlos & L. Ingber (2003) A reinforcement learning method based on adaptive simulated annealing, In: Proceedings International Midwest Symposium on Circuits and Systems (MWCAS), December 2003, IEEE CAS, 121-124. [URL https://www.ingber.com/asa03_reinforce.pdf]

I. Bray, A. Gibson & J. White (2020) Coronavirus disease 2019 mortality: a multivariate ecological analysis in relation to ethnicity, population density, obesity, deprivation and pollution. Public Health. 185, 261-263. [URL https://doi.org/10.1016/j.puhe.2020.06.056]

L. Ingber (1993) Statistical mechanics of combat and extensions, In: Toward a Science of Command, Control, and Communications, ed. C. Jones. American Institute of Aeronautics and Astronautics, 117-149. [ISBN 1-56347-068-3. URL https://www.ingber.com/combat93_c3sci.pdf]

L. Ingber (1996) Canonical momenta indicators of financial markets and neocortical EEG, In: Progress in Neural Information Processing, ed. S.-I. Amari, L. Xu, I. King & K.-S. Leung. Springer, 777-784. [Invited paper to the 1996 International Conference on Neural Information Processing (ICONIP'96), Hong Kong, 24-27 September 1996. ISBN 981-3083-05-0. URL https://www.ingber.com/markets96_momenta.pdf]

L. Ingber (2010) Trading in Risk Dimensions, In: The Handbook of Trading: Strategies for Navigating and Profiting from Currency, Bond, and Stock Markets, ed. G.N. Gregoriou. McGraw-Hill, 287-300.

L. Ingber (2012) Adaptive Simulated Annealing, In: Stochastic global optimization and its applications with fuzzy adaptive simulated annealing, ed. H.A. Oliveira, Jr., A. Petraglia, L. Ingber, M.A.S. Machado & M.R. Petraglia. Springer, 33-61. [Invited Paper. URL https://www.ingber.com/asa11_options.pdf]

L. Ingber (2015) Biological Impact on Military Intelligence: Application or Metaphor?. International Journal of Intelligent Defence Support Systems. 5(3), 173-185. [URL https://www.ingber.com/combat15_milint.pdf]

L. Ingber (2018) Quantum calcium-ion interactions with EEG. Sci. 1(7), 1-21. [URL https://www.ingber.com/smni18_quantumCaEEG.pdf and https://doi.org/10.3390/sci1010020]

L. Ingber (2020a) Developing bid-ask probabilities for high-frequency trading. Virtual Economics. 3(2), 7-24. [URL https://www.ingber.com/markets19_bid_ask_prob.pdf and https://doi.org/10.34021/ve.2020.03.02(1)]

L. Ingber (2020b) Forecasting with importance-sampling and path-integrals: Applications to COVID-19. Report 2020:FISPI. Physical Studies Institute. [URL https://www.ingber.com/asa20_forecast.pdf and https://doi.org/10.20944/preprints202009.0385.v3]

L. Ingber (2020c) Quantum calcium-ion affective influences measured by EEG. Report 2020:QCIA. Physical Studies Institute. [URL https://www.ingber.com/quantum20_affective.pdf and https://doi.org/10.20944/preprints202009.0591.v1]

L. Ingber & R.P. Mondescu (2001) Optimization of trading physics models of markets. IEEE Transactions Neural Networks. 12(4), 776-790. [Invited paper for special issue on Neural Networks in Financial Engineering. URL https://www.ingber.com/markets01_optim_trading.pdf]

L. Ingber & R.P. Mondescu (2003) Automated internet trading based on optimized physics models of markets, In: Intelligent Internet-Based Information Processing Systems, ed. R.J. Howlett, N.S. Ichalkaranje, L.C. Jain & G. Tonfoni. World Scientific, 305-356. [Invited paper. URL https://www.ingber.com/markets03_automated.pdf]

A.Y. Li, T.C. Hannah, J. Durbin, N. Dreher, F.M. McAuley, N. F. Marayati, Z. Spiera, M. Ali, A. Gometz, J.T. Kostman & T.F. Choudhri (2020) Multivariate analysis of black race and environmental temperature on COVID-19 in the US. The American Journal of the Medical Sciences. 360(4), 348-356. [URL https://doi.org/10.1016/j.amjms.2020.06.015]

C. Okoli & S.D. Pawlowski (2004) The Delphi method as a research tool: an example, design considerations and applications. Information and Management. 42(1), 15-29. [URL https://doi.org/10.1016/j.im.2003.11.002]