Newsjournal of the Society for Industrial and Applied Mathematics sinews.siam.org

siam news

Volume 54/ Issue 4
May 2021

Wave-Structure Interactions and Wave Energies

By David Lannes

"*Consider the action of the waves — the ebb and flow of the tides. What is the ocean? A prodigious force wasted.*" Back in 1874, when French writer Victor Hugo published his novel *Ninety-Three*, mankind already knew that the ocean could provide energy. Yet the first industrial exploitation of ocean energy did not come to light until 1967—nearly a century later—at the Rance Tidal Power Station in France's Brittany region. A revived interest in marine renewable energies arose in the 2000s, inspired by an oil shock and the need to increase renewable energy use due to the carbon emissions of fossil fuels. Technological and computational advances also made renewable energies more realistic.

Renewable marine energies include tidal energy, offshore wind energy, ocean currents, waves, thermal potential, salinity gradient power, and even biomass. The most widely spread of these energies—and the one with the most potential—is offshore wave energy, which has a technical potential of roughly 16,000 terawatt hours $(\mathrm{TW\cdot h})$ per year by 2050 [2]. The technology is the same for onshore wind turbines, but additional engineering challenges arise — such as the construction of offshore foundations or floating support structures (in deeper water). To design the latter, one must understand the interactions between waves and floating structures.

Wave-structure interactions are also obviously important in the concept of wave energy. With an estimated technical potential of 5,600 $\mathrm{TW\cdot h}$ per year, wave energy is less powerful than offshore wind energy but still offers interesting perspectives. Researchers have proposed several devices to exploit wave energy; the Pelamis Wave Energy Converter in the U.K., which consists of several partially submerged connected cylinders whose relative motion generates electricity when a wave passes, was the first to actually generate electricity into the grid. Since then, scientists have developed dozens of other systems, including floating buoys, oscillating water columns, and flaps.

To design floaters, offshore industry has traditionally employed tank testing at small scales. However, approaches that are based on computational fluid dynamics (CFD) are beginning to convincingly replace model tests, especially for the study of turbulence-related issues. CFD computations are particularly relevant in the analysis of wave impact and vortex-induced motion that can endanger the mooring systems of floating structures, among other applications.

Although CFD is a common tool in aeronautics and the automotive industry, it remains at an early stage of development for floater design. A recent study [5] identified the following four unique challenges that offshore engineers face that differ from other industries:

(1) A highly separated flow with a Reynolds number of order 10^7 around the floater hull

(2) Large scale differences between the hull and the mooring and riser systems

(3) An open ocean environment that requires the computation of a large volume of fluid

(4) A non-Gaussian stochastic environment that necessitates multiple computations to provide reliable statistics for floater motion.

Due to these difficulties, the estimated cost of a CFD project for a numerical basin is equivalent to that of physical model tests.

An interesting alternative to CFD computations involves describing the flow via the so-called fully nonlinear potential flow (FNPF) approach [8], during which the flow is assumed to be irrotational and the velocity field derives from a scalar velocity potential; one can find the latter in the fluid domain by solving a Laplace equation with appropriate boundary conditions on the free surface and the object's sides. Although this approach cannot account for turbulence-related issues, it can capture nonlinear effects in the wave-structure interactions.

See **Wave Energies** *on page 3*

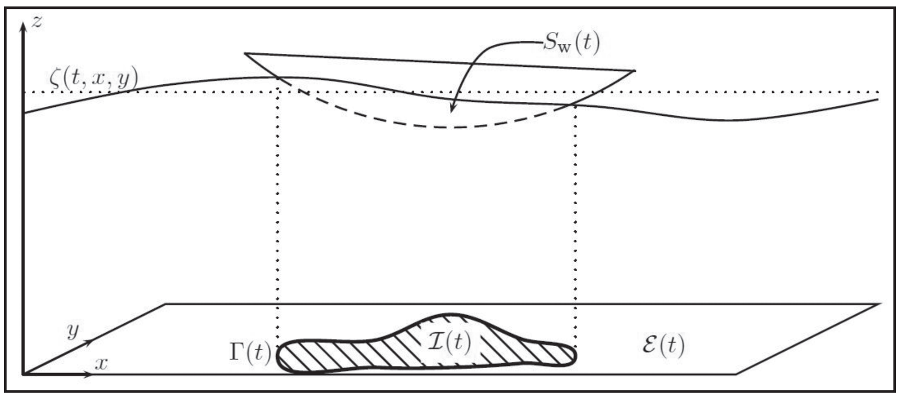


Figure 1. *Interior region $\mathcal{I}(t)$, exterior region $\mathcal{E}(t)$, and projection of the contact line $\Gamma(t)$. Figure adapted from [6].*

Ecological Transients and the Ghost of Equilibrium Past

By Matthew R. Francis

The sight and smell of eutrophication—in the form of a layer of stinking green algae on a lake or pond—is likely familiar to many readers. The result is detrimental, even toxic, to other species that rely on the water, ranging from tiny animals to birds and even humans. For example, eutrophication on Lake Erie affects millions (see Figure 1). But the real culprit is actually the substance that feeds the algae: excess phosphorous that is produced by human activities like fertilizer runoff and leaky septic systems.

To manage eutrophication, one must know whether the affected body of water resides in a eutrophic stable state, or if its state is a *long transient*. The second case mimics stability because it can last a long time but is sustained by another source of phosphorous in the lakebed sediments. According to Tessa Francis, an ecologist at the University of Washington Puget Sound Institute, the wrong management choice has major consequences in terms of costs and trade-offs.

"You're investing all of this social, political, and economic capital into management, but you're getting no results from it," Francis said. "If you gave the system a bigger smack by adding an alternative management strategy to tackle the phosphorus pool at the bottom of the lake, that would be more likely to get your lake back to the state you want. This is just one consequence of long transients in terms of how they affect management decisions."

Mathematical biologists, field ecologists, and other researchers study the dynamics of ecological long transients to understand how to best distinguish them from stable or oscillatory states. This theme was the topic of a session at the virtual 2021 American Association for the Advancement of Science Annual Meeting[1] this February.

"Transient dynamics are everything that you see when a system is away from equilibrium," Karen Abbott, a theoretical ecologist at Case Western Reserve University, said. "It's either in the process of approaching equilibrium, or permanently disturbed in a way that will never allow [the system] to fully approach equilibrium."

The Ghost in the Saddle

Dynamical systems research has historically focused on the identification of asymptotic behaviors, including stable mathematical attractors, regular oscillations, and other limit cycles. But as Abbott asserts, transients (rather than stability) may be the norm in real ecological systems. "If we're going to use observations of the natural world to discover the rules through which ecosystems are structured, we need to understand that what we're looking at is probably not in equilibrium," she said. Between internal fluctuations and external influences like weather or human intervention, ecological systems balance on metastable mathematical ledges that are far from the clean asymptotic behaviors that dynamicists prefer.

To further complicate the situation, these transients can last a long time and mimic asymptotic conditions. In fact, it is difficult for researchers to precisely define "long." "In ecological applications, a system usually has a certain characteristic time," Sergei Petrovskii of the University of Leicester said. This timescale can encompass the life span of an individual or multiple generations of a species in a habitat. "A common-sense approach is to call 'long' anything that will substantially exceed this characteristic time by, say, an order of magnitude," he continued. "As mathematicians, we tend to call a transient 'long' when it can be made infinitely long if we play with a certain parameter."

Petrovskii is concerned with the classification of the mathematical behavior of long transients. "Ghost attractors," which are asymptotic steady-state solutions under slightly different values of the controlling

See **Ecological Transients** *on page 4*

Figure 1. *A satellite photo of Lake Erie in North America (taken in October 2011) shows toxic algae in green. The spread pattern is due primarily to phosphorus runoff from agriculture. Public domain image.*

---

[1] https://www.aaas.org/events/2021-aaas-annual-meeting

Nonprofit Org
U.S. Postage
PAID
Permit No 360
Bellmawr, NJ

siam®
SOCIETY for INDUSTRIAL and APPLIED MATHEMATICS
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA

siam news
Volume 54/ Issue 4/ May 2021

Renewables Reliability in an Era of Force Majeure

By René Aïd, Mike Ludkovski, and Ronnie Sircar

Following the global lockdown in March 2020, the SIAM Activity Group on Financial Mathematics and Engineering[1] launched a virtual seminar series[2] to maintain a sense of connectivity within the worldwide financial mathematics community. In addition to individual talks, the series included an inaugural panel discussion about energy markets that focused on the COVID-19 pandemic's effect on energy production, novel negative oil prices, and research opportunities from renewable sources' increased impact on electricity production. The panel, which was moderated by Ronnie Sircar (Princeton University), consisted of René Aïd (Paris Dauphine University), Glen Swindle (Scoville Risk Partners), Zef Lokhandwalla (Bloomberg L.P.), and Mike Ludkovski (University of California, Santa Barbara).

The horrific power outages (and ensuing consequences) in Texas during February 2021 served as a prime example of the ramifications of electric grids that are run as essentially deterministic systems; they are unprepared for extreme weather events, despite the introduction of more weather-dependent production sources. A decade ago, experts were concerned with electricity price spikes in Texas due to summer heat waves [7]. Now the southern U.S. states must determine whether recent events will inspire them to winterize their generation units — and if so, to what degree and at what cost.

Aside from upending our daily lives, COVID-19 has dramatically impacted the energy industry. Lockdowns both significantly lowered overall energy consumption and changed the nature of this consumption. For example, the shift to a work-from-home lifestyle and erosion of the typical 9 a.m. to 5 p.m. pattern placed strains on electricity grid operators, who had to simultaneously adjust to daily time shifts and spatial changes in energy consumption (from commercial cores to residential neighborhoods). The resulting change in the shape of electricity demand will likely be long lasting (see Figure 1).

As another COVID-19-driven effect, April 2020 saw one of the most spectacular examples of a price collapse. Crude oil prices dropped into negative territory, touching a low of -$40.32 on April 20, 2020 [4]. While financial prices tend to be positive, the need to actually take delivery and then store all that crude—a physical commodity—created giant market distortions. When every available storage facility filled up (due to decreased demand) and significant bottlenecks arose, financial traders were forced to *pay* to get rid of their crude futures and respective obligations to receive physical barrels. The result was a several-day trading extravaganza, with reports of lost and gained fortunes and even some auxiliary systems shutting down, as they were not programmed to expect negative prices.

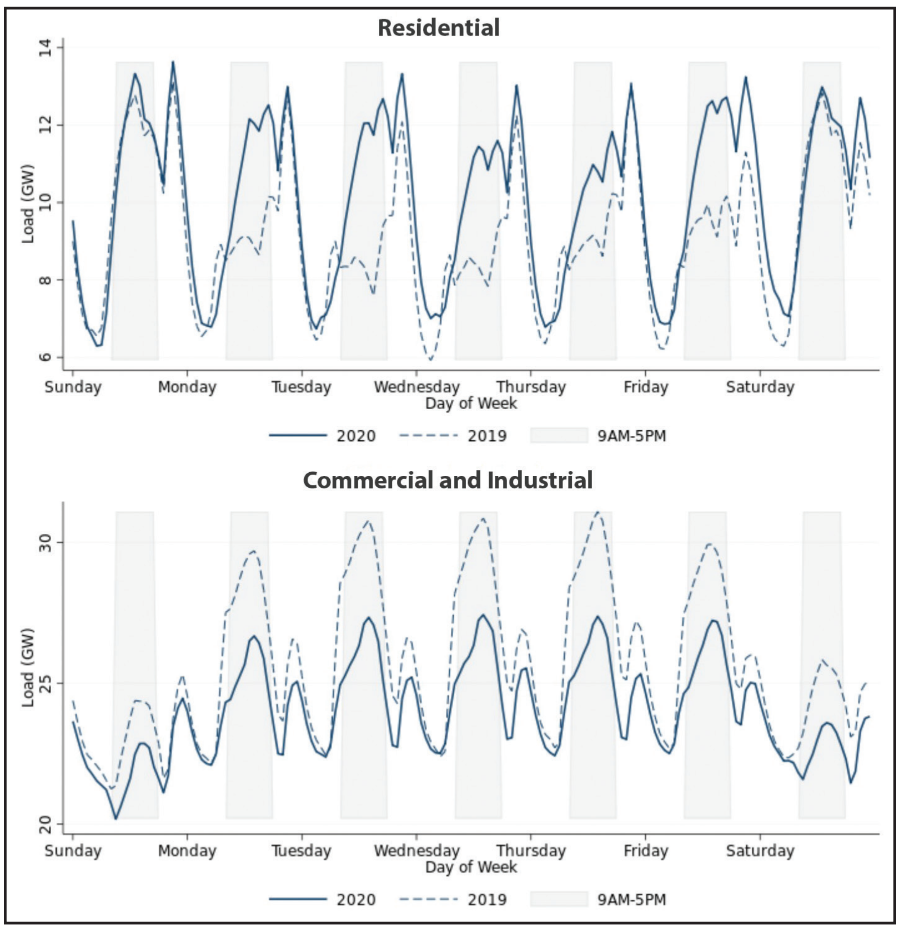


Figure 1. *Temperature-adjusted electricity consumption in Texas by customer class for April/May 2020 versus the same period in 2019. Figure courtesy of Steve Cicala [6].*

In a related development, the pandemic triggered the little-known force majeure clause of electricity futures contracts in France. Electricity suppliers invoked the clause to suspend their obligation to buy electricity from the national Electricité de France (EDF). EDF operates the French nuclear plants and is regulated to sell between 100 and 150 terawatt hours of nuclear energy via a forward contract at the fixed 42 euro/megawatt hour price. Retailers bought all of this power forward in October 2019; but by March 2020, when they should have taken delivery, the spot electricity price had crashed — along with the demand. To avoid large financial losses, the buyers claimed force majeure. The case was initially settled in their favor but is currently under appeal.

In contrast to crude oil, negative electricity prices have become commonplace all over the world. The culprit is the sun; power demand remains steady during midday, but solar production is now enormous — for example, it supplies more than 50 percent of total demand in California. As a result, there is really *too much* solar energy in some places at certain times. To keep the lights on in the evening, operators regularly curtail solar (and often wind) energy generation so that fossil-fueled plants with slow ramping times can remain active and economically viable (see Figure 2, on page 5). At the same time, renewable generation is highly noisy and can experience forecast errors of up to 10 percent, even on a day-ahead time scale; large deviations cause price spikes.

Tackling this challenge requires collaboration between applied mathematicians, power system engineers, and policymakers. It starts with quantification of the uncertainties in daily grid operation and eventually will necessitate a redesign of the electricity markets that remain geared towards *firm* energy suppliers. Decentralization, improved forecasting, faster optimization of distribution and dispatch, and risk allocation are all imperative for a smooth transition to a renewables-based grid. Recently, the Advanced Research Project Agency–Energy (ARPA-E) within the U.S. Department of Energy ran the Performance-based Energy Resource Feedback, Optimization, and Risk Management (PERFORM) solicitation [1] to address this challenge; it is now funding 13 multidisciplinary projects. Application of the mathematical toolbox to manage renewables uncertainty is an emerging theme. The new paradigm requires researchers to embrace the unavoidable fluctuation of wind and solar energy availability, and re-engineer our energy systems accordingly. *SIAM News* readers can contribute to this enterprise by engaging with climate scientists to enhance numerical weather forecasting systems for probabilistic location-specific solar and wind predictions, or working with power engineers to build faster solvers for the highly non-convex unit commitment problems that are necessary for hour-by-hour balancing.

Thanks to ever more cost-effective technology, demand for renewables grew unabatedly even as overall energy consumption decreased during lockdown. 2020 therefore shattered many records; renewable energy even surpassed coal production in the Electric Reliability Council of Texas grid. One component of this shift is the *decentralization* of energy production that is exemplified by the growing use of rooftop solar panels. Such behind-the-meter resources—which are managed by consumers rather than utility companies—open a new frontier in energy use management and

See **Renewables Reliability** *on page 5*

---

[1] https://www.siam.org/membership/activity-groups/detail/financial-mathematics-and-engineering

[2] http://wiki.siam.org/siag-fm/index.php/Current_events

siam news

ISSN 1557–9573. Copyright 2021, all rights reserved, by the Society for Industrial and Applied Mathematics, SIAM, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688; (215) 382-9800; siam@siam.org. To be published 10 times in 2021: January/February, March, April, May, June, July/August, September, October, November, and December. The material published herein is not endorsed by SIAM, nor is it intended to reflect SIAM's opinion. The editors reserve the right to select and edit all material submitted for publication.

Advertisers: For display advertising rates and information, contact Kristin O'Neill at marketing@siam.org.

One-year subscription (nonmembers): Electronic-only subscription is free. $73.00 subscription rate worldwide for print copies. SIAM members and subscribers should allow eight weeks for an address change to be effected. Change of address notice should include old and new addresses with zip codes. Please request address change only if it will last six months or more.

Printed in the USA.
siam *is a registered trademark.*

Call for Nominations for 2023 ICIAM Prizes

The International Council for Industrial and Applied Mathematics (ICIAM) Prize Committee calls for nominations for the six ICIAM prizes to be awarded in 2023: the Collatz Prize, the Lagrange Prize, the Maxwell Prize, the Pioneer Prize, the Su Buchin Prize, and the Industry Prize. Each ICIAM prize has its own unique attributes but all are truly international in nature; nominations are therefore welcome from every part of the world. A nomination should account for the specifications of a particular prize[1] and contain the following information:

- Full name and address of the nominee
- Nominee's webpage, if applicable
- Name of particular ICIAM prize
- Justification for nomination; in particular, nominators must cite their reasons for selecting the candidate, including explanations of the scientific and practical influence of the candidate's work and publications
- Proposed citation: a concise statement about the nominee's outstanding contributions in less than 250 words
- CV of the nominee
- Two to three letters of support from experts in the field, and/or two to three names of experts that the ICIAM Prize Committee can consult
- Name and contact details of the nominator.

Nominations should be made electronically through the ICIAM Prizes website.[2] The deadline for nominations is **September 1st, 2021**. Please contact **president@iciam.org** with any questions about the nomination procedure.

Prize Descriptions

The **Collatz Prize** provides international recognition to individual scientists who are under 42 years of age for outstanding work in industrial and applied mathematics. A recipient's 42nd birthday must not occur before January 1 of the year in which the prize is presented.

The **Lagrange Prize** acknowledges individual mathematicians who have made exceptional contributions to applied mathematics throughout their careers.

The **Maxwell Prize** offers international recognition to a mathematician who has demonstrated originality in applied mathematics.

The **Pioneer Prize** recognizes pioneering work that introduces applied mathematical methods and scientific computing techniques to an industrial problem area or new scientific field of applications.

The **Su Buchin Prize** provides international recognition for an outstanding contribution by an individual in the application of mathematics to emerging economies and human development, particularly at the economic and cultural level in developing countries. This contribution can include efforts to improve mathematical research and teaching in those countries.

The **Industry Prize** honors scientists who have made outstanding contributions to innovative mathematical techniques with a demonstrated impact in industry.

ICIAM Prize Committee for 2023

- Ya-xiang Yuan, Committee Chair
- Barbara Wohlmuth, Chair of the Collatz Prize Subcommittee
- Leah Edelstein-Keshet, Chair of the Lagrange Prize Subcommittee
- Gang Bao, Chair of the Maxwell Prize Subcommittee
- Alfredo Bermudez, Chair of the Pioneer Prize Subcommittee
- Lois Curfman McInnes, Chair of the Su Buchin Prize Subcommittee
- Nira Chamberlain, Chair of the Industry Prize Subcommittee.

ICIAM is the world organization for applied and industrial mathematics. Its members consist of mathematical societies that are based in more than 30 countries. See the ICIAM website[3] for more information.

---

[1] http://www.iciam.org/iciam-prizes

[2] https://iciamprizes.org

[3] http://www.iciam.org

Wave Energies

Continued from page 1

The performance of CFD and FNPF computations is steadily improving, but engineers use much simpler methods to construct "wave-to-wire" models that simulate the interactions of the various devices of arrays that count up to dozens of wave energy converters, as well as their connection to the network [4]. Such methods are typically based on the Cummins' equations [3]: a set of coupled linear integro-differential equations, one for each degree of freedom. In the absence of an exciting force, the equations take the form

$$\sum_{j=1}^{6}\Big[(m_j\delta_{jk}+m_{jk})\ddot{x}_j+c_{jk}x_j+ \int_{-\infty}^{t}K_{jk}(t-\tau)\dot{x}_j(\tau)\mathrm{d}\tau\Big]=0 \quad (1)$$
$$(1\leq k\leq 6).$$

Here, m_j denotes the mass/inertia in the jth degree of freedom, δ_{jk} is the Kronecker symbol, and $c_{jk}x_j$ is the hydrostatic force in the kth mode that results from a perturbation x_j in the jth mode. Finally, $(m_{jk})_{jk}$ and $(K_{jk})_{jk}$ are the matrices of added mass and radiation impulse response functions, respectively. In order to define these latter quantities and understand the limitations of the model, one can sketch the model's derivation.

We assume that the flow is linear and irrotational, and that the surface elevation's variations are neglected for the velocity potential's domain of definition. One can decompose the velocity potential Φ, which results from a small perturbation $x_j(t)$ of the floating object in the jth mode, into

$$\Phi=\dot{x}_j\psi_j+\int_{-\infty}^{t}\varphi_j(t-\tau)\dot{x}_j(\tau)\mathrm{d}\tau. \quad (2)$$

Here, ψ_j is the potential directly associated with the instantaneous impulsive velocity of the floating object, and φ_j is the potential associated with the radiating disturbance of the free surface (the time integral accounts for disturbances of the surface that are created by the object's previous displacements). On the other hand, the pressure P at the wetted surface S_{w} of the object (whose time variations are neglected) is given by the linear approximation of Bernoulli's equation:

$$P=-\rho gz-\rho\partial_t\Phi \quad \text{on } S \qquad (3)$$

(ρ is the constant density of the fluid, g is the gravity, and z is the vertical variable). One can therefore write the resulting force that is exerted on the solid as

$$F=\int_{S_{\mathrm{w}}}\Big(-\rho gz-(\partial_t\Phi)\Big)\mathbf{n} \qquad (4)$$

($\mathbf{n}$ is the outward normal to the fluid surface under the object). A similar expression holds for the resulting torque. Using (2)-(3), Newton's equations, and Archimedes' principle, one readily obtains (1) with

$$m_{jk}=\rho\int_{S_{\mathrm{w}}}\psi_j s_k \quad \text{and}$$
$$K_{jk}(\tau)=\rho\int_{S_{\mathrm{w}}}\partial_t\varphi_j(\tau)s_k.$$

Here, $s_k=\mathbf{n}\cdot\mathbf{e}_k$ $(k=1,2,3)$ or $s_k=(\mathbf{r}\times\mathbf{n})\cdot\mathbf{e}_{k-3}$ $(k=4,5,6)$, where $\mathbf{e}_k$ is the unit vector in the kth direction and $\mathbf{r}$ is the position vector with respect to the object's center of gravity. Engineers then compute the various components of the velocity potential with commercial software programs like WAMIT, which are based on a representation of the potential in terms of Green's functions.

The assumption of linearity is ubiquitous in the derivation of Cummins' equations (1): the variations of the immersed part S_{w} of the body are neglected, wave motion is linear, nonlinear terms are neglected in Bernoulli's equation, and so forth. Models based on Cummins' equations that have proven very useful for the study of floating structures are therefore not utilized when nonlinear effects are important, i.e., if one wants to assess maximum loads on the structure in extreme sea conditions. Such a situation is likely to occur quite often in shallow water, where the waves are of larger amplitude.

To mitigate this scenario, I have recently proposed another approach [6] that somehow fits between linear approaches that are based on Cummins' equations and the very precise but computationally expensive CFD (and even FNPF) approaches. This method is essentially based on three steps:

(1) Use an asymptotic model to describe the waves

(2) Consider the pressure that is exerted on the object as a Lagrange multiplier

(3) Reduce the problem to a transmission problem.

Regarding the first step, scientists have made considerable progress in the last two decades on the derivation and justification of asymptotic models to the $d+1$-dimensional free surface Euler equations (also called water waves equations), where d is the horizontal dimension — especially in shallow water [7]. Such models typically couple the elevation ζ of the free surface with respect to the rest state with the horizontal discharge Q (the vertical integral of the horizontal velocity). An obvious simplification involves casting the equations on the horizontal domain (with no dependence on the vertical variable z). Both quantities are connected by the exact relation

$$\partial_t\zeta+\mathrm{div}_{\mathrm{h}}Q=0,$$

where $\mathrm{div}_{\mathrm{h}}$ is the divergence with respect to the horizontal variables. This equation is complemented by an approximate evolution equation for Q, which takes the form

$$\begin{aligned}\partial_t Q+g(h_0+\zeta)\nabla_{\mathrm{h}}\zeta+&\\ \ldots=-\frac{1}{\rho}(h_0+\zeta)\nabla_{\mathrm{h}}P_{\mathrm{surf}}.&\end{aligned} \qquad (5)$$

Here, h_0 is the water depth at rest and P_{surf} is the pressure at the surface; the dots in the equation depend on the range of validity and precision of the asymptotic model.

For the second step, one can divide the horizontal plane into three parts: the interior region (the projection of the object's time-dependent wetted surface $S_{\mathrm{w}}(t)$), the projection of the contact line, and the exterior region (see Figure 1, on page 1). The surface in the *exterior* region is free but the pressure is constrained; P_{surf} is typically equal to a constant atmospheric pressure so that the right side vanishes in (5). The reverse occurs in the *interior* region: the pressure is free but the fluid's surface is constrained and must coincide with the wetted part of the object. One can interpret the right side in (5), which does not vanish, as the Lagrange multiplier that is associated with this constraint.

The third step involves finding the pressure for the incompressible Euler equations in the interior domain by solving a rather simple elliptic equation. When combined with some coupling conditions at the contact line, this solution allows one to reduce the whole problem to a transmission problem in the exterior region for the asymptotic model, with nonstandard transmission conditions.

This three-step approach is quite simple and very efficient. It accounts for nonlinear effects and benefits from the recent progress with numerical simulations of shallow water flows. The method also yields several nonlinear generalizations of Cummins' equations and shows that researchers should generally not expect to describe a floater's motion via simple integro-differential equations; they must instead consider transmission problems. The resulting models are mathematically challenging. Most importantly, one should carefully investigate this approach's range of validity by comparing it with CFD computations or numerical approaches that are fully based on nonlinear flows or experiments.

Ultimately, wave structure comprises only a small part of the problems that are associated with the development of marine renewable energies. For example, seawater is corrosive and warrants the creation of specific materials; precise weather and wave forecasts are needed; and researchers must evaluate wave farms' effect on the wave field, in addition to the impact of sedimentation, erosion, and water quality on biodiversity, local fisheries, and so on. Seemingly more remote issues—such as legal aspects, funding policies, economic viability, and social acceptance of altered landscapes—are also relevant.

Although marine renewable energies may be "greener" than fossil energies, they are certainly not "green." As scientists, we can help to fairly evaluate the merits and disadvantages that will allow citizens to make informed choices about the society in which they want to live. Some communities have already created scientific structures to assist local governments that are deciding on strategies to address global warming [1]. Answering such concerns in turn raises new scientific issues; this is a new and fascinating form of interdisciplinarity.

This article is based on David Lannes' invited presentation at the 2020 SIAM Annual Meeting,[1] which took place virtually last year.

References

[1] AcclimaTerra, & Le Treut, H. (Ed). (2020). *Anticipating climate change in Nouvelle-Aquitaine. To guide policy at local level* (Executive Report). Éditions AcclimaTerra.

[2] Borthwick, A.G. (2016). Marine renewable energy seascape. *Engineer.*, *2*(1), 69-78.

[3] Cummins, W. (1962). *The impulse response function and ship motions.* (Hydromechanics Laboratory Research and Development Report). U.S. Department of the Navy.

[4] Forehand, D.I., Kiprakis, A.E., Nambiar, A.J., & Wallace, A.R. (2015). A fully coupled wave-to-wire model of an array of wave energy converters. *IEEE Trans. Sustain. Ener.*, *7*(1), 118-128.

[5] Kim, J.W., Jang, H., Baquet, A., O'Sullivan, J., Lee, S., Kim, B., …, Jasak, H. (2016). Technical and economic readiness review of CFD-based numerical wave basin for offshore floater design. In *Offshore Technology Conference*. Houston, TX.

[6] Lannes, D. (2017). On the dynamics of floating structures. *Annals of PDE*, *3*(1), 11.

[7] Lannes, D. (2020). Modeling shallow water waves. *Nonlinearity*, *33*, R1.

[8] Penalba, M., Giorgi, G., & Ringwood, J.V. (2017). Mathematical modelling of wave energy converters: A review of nonlinear approaches. *Renew. Sustain. Ener. Rev.*, *78*, 1188-1207.

David Lannes is a CNRS research director at the Institut de Mathématiques de Bordeaux. He works on the modelling and mathematical analysis of waves in coastal regions, and is interested in applications like the evaluation of submersion risks or renewable energies in marine environments. Lannes is also involved in various interdisciplinary structures that aim to coordinate the scientific community's answers to ongoing environmental issues.

---

[1] https://www.siam.org/conferences/cm/conference/an20

Ecological Transients

Continued from page 1

parameter, are one important class (see Figure 2a). But when the ghost is present, the system is attracted into the long transient and lingers there for an extended period of time before moving on to a real equilibrium state [3].

A second case occurs when the system contains a saddle point: a state that resembles an attractor in one direction but is not at stable equilibrium. If the slope away from the saddle is shallow, the system lingers near the unstable equilibrium, mimicking stability and producing the long transient behavior known as a crawl-by (see Figure 2b). This behavior occurs in systems with multiple timescales—such as interacting species with different life spans—or large stochastic effects.

Abbott is particularly interested in stochasticity's role in long transients. In the ubiquitous Lotka-Volterra model for competing species, for instance, the coupled equations are deterministic and cannot account for many important ecological parameters.

"Stochastic disturbances jostle the system to other combinations of population sizes or other states," Abbott said. "Lots of different species are ectotherms, so they respond directly to variations in temperature. Things like feeding rates and birth rates—very fundamental demographic properties—are going to respond to temperature."

To demonstrate this phenomenon, Abbott and her colleagues began with the Lotka-Volterra model for the population densities of two species (u_1, u_2) in the form

$$\frac{du_1}{dt} = r_1 u_1 (1 - u_1 - \rho\alpha u_2) + \varepsilon_1$$

$$\frac{du_2}{dt} = r_2 u_2 (1 - u_2 - \alpha u_1) + \varepsilon_2.$$

Here, r_k are the natural growth rates, α and ρ are interaction parameters, and ε_k is Gaussian noise with a mean of 0 and variance of σ [1]. The group considered a case wherein species 1 is native and species 2 is invasive, so that $r_2 > r_1$ and $\rho > 1$ to give the invaders a slight advantage. In the absence of noise, this model has an equilibrium point at $(u_1, u_2) = (1, 0)$ that represents a high density of the native species and (local) extinction of the invasive population.

If the relative amount of competition ρ is fixed but α varies, the system has two bifurcations in the absence of noise. When $\alpha < 1/\rho$, neither species has a competitive edge and they coexist; when $1/\rho < \alpha < 1$, the system favors the invaders at the expense of the endemic species; and when $\alpha > 1$, either species can "win" the competition, thus producing a bistable state. Including noise in the equations changes the equilibrium state to a saddle point, so that local extinction of the invasive species is a long transient and stochasticity allows the invaders to return. This incidence has a profound effect on potential management decisions, since adding native species alone is not enough; active removal of invasive species or other mitigation options must also occur [2].

"We tend to monitor ecosystems on relatively short timescales and assume that if what we're observing looks like equilibrium, it therefore is at equilibrium," Francis said. "If you don't have long-term monitoring of ecosystems, then you're unable to see the sort of long-term dynamics that play out. This may affect your ability to adequately manage the system."

Management and Modeling

Although additive white noise is of course not a good model for realistic ecological systems, it has the virtue of simplicity. After all (to paraphrase the first sentence of Leo Tolstoy's *Anna Karenina*), Gaussian noise is all alike, but each non-Gaussian noise type is non-Gaussian in its own way.

"The longer-term goal is to become more sophisticated and begin asking, 'How do we think these systems are really being perturbed in nature?'" Abbott said. "If a storm comes through and kills a lot of individuals, do we want to represent that as a random spike in the death rate? Or do we want to randomly perturb the whole population density downward? These are two ways to represent the same thing, but they don't always yield the same outcome mathematically."

Another challenge stems from the fact that real ecosystems are more complex than even the most sophisticated dynamical models. "Cases are few and far between where you can actually match a model of a long transient to a real-world example," Francis said. "In some cases, we can use these models to understand the controlling parameters on long transients."

Instead, much of modeling's value lies in the qualitative understanding of system behaviors. Without this mathematical insight, it is difficult for researchers to adjust intervention strategies for invasive species or identify why eutrophic lakes resist mitigation.

Current work on long transients mirrors early research on nonlinear dynamics, when scientists devoted much effort to identifying, cataloging, and developing techniques to better comprehend complex systems. Because ecosystems are intrinsically out of equilibrium, we must adjust our thinking about stability and timescales to match the world we live in — and change how we interact with it.

References

[1] Abbott, K.C., & Dakos, V. (2020). Mapping the distinct origins of bimodality in a classic model with alternative stable states. *Theor. Ecol.*

[2] Francis, T.B., Abbott, K.C., Cuddington, K., Gellner, G., Hastings, A., Lai, Y.-C., …, Zeeman, M.L. (2021). Management implications of long transients in ecological systems. *Nat. Ecol. Evol.*, *5*(3), 285-294.

[3] Hastings, A., Abbott, K.C., Cuddington, K., Francis, T., Gellner, G., Lai, Y.-C., …, Zeeman, M.L. (2018). Transient phenomena in ecology. *Science*, *361*(6406), eaat6412.

Matthew R. Francis is a physicist, science writer, public speaker, educator, and frequent wearer of jaunty hats. His website is BowlerHatScience.org.

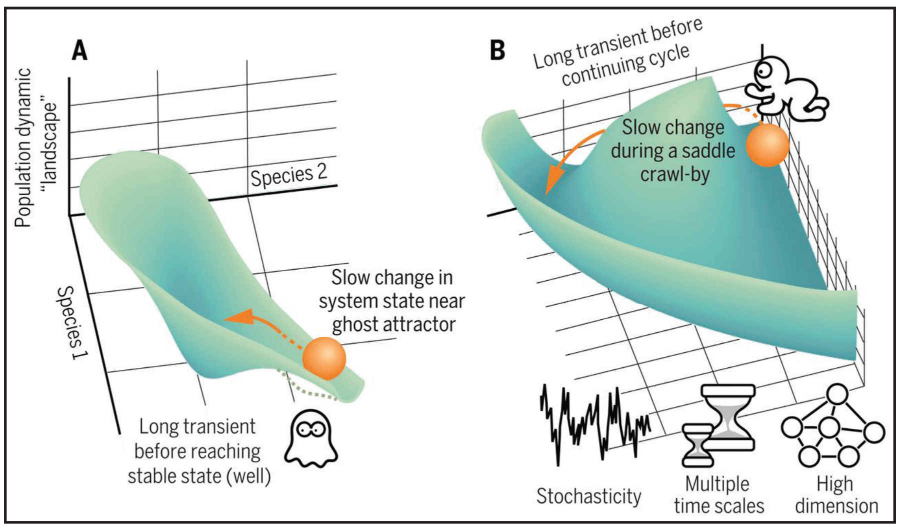


Figure 2. *Visualization of two major mathematical types of long transients.* **2a.** *A "ghost attractor" would be a stable equilibrium point under different conditions, but the system eventually moves away from the ghost because the mathematical shape of the transient contains an escape.* **2b.** *A "crawl-by" is a saddle point in the dynamical space, meaning that it is attractive in one direction and repulsive in the other. Because it is nearly flat, it mimics stable equilibrium. Image courtesy of [3]. Reprinted with permission from AAAS.*

When Software Harms, What You Reap Is What You Sow

By Matthew R. Francis

As of July 2020, the CalGang database[1] contained the names and personal details of nearly 90,000 people in the state of California who are suspected of being in gangs or associating with gang members. Despite its stated purpose to provide law enforcement agencies with accurate intelligence information, audits and independent investigations revealed that the database was riddled with errors, falsified material, racial profiling, and other serious problems.

Databases and algorithms are ubiquitous parts of our interconnected world, but CalGang illustrates a major way in which they can fail people. If a streaming service suggests a movie that you do not like, no real harm is done; but if your name appears in CalGang, you may face consequences like increased police harassment or harsher sentences if charged with a crime.

"[Most of] the people creating these technologies are not affected in negative ways," Seny Kamara, a computer scientist at Brown University, said. "But if you're a young Black male growing up in Chicago or New York or California, you know that you may end up as a false positive in a gang database, and that affects your life in a completely different way."

During his presentation at the 2021 American Association for the Advancement of Science Annual Meeting,[2] which took place virtually this February, Kamara used CalGang as a specific example of the disconnect between people who design software and those who are harmed by it. Although these databases are frequently promoted as more objective than human decision making, they often perpetuate the same prejudices and harms.

"One of the rationales for risk assessment tools [in criminal justice] is that judges can be inconsistent or make decisions that are not always suitable," Suresh Venkatasubramanian, a computer scientist at the University of Utah who also spoke at the AAAS session, said. "An algorithmic process might help us, [but] then there is a concern that these systems—by merely looking at prior data—are amplifying patterns of bias, especially racial bias. The question then becomes, 'Why are we building a risk assessment tool in the first place?' These are always being commissioned and built by the folks who are putting people in jail."

In other words, improving software like CalGang still accepts the necessity of its existence, and the technologists who create it are complicit in the abuses of the carceral system — regardless of their own intentions. Both Kamara and Venkatasubramanian indicate that computer scientists often choose a side by merely accepting work that is commissioned by the police, military, or private companies. Though taking a position is not intrinsically bad, the researchers agree that one must understand the ethics and consequences that are associated with building algorithms, databases, and other software tools.

On the surface, CalGang—and other applications like it—may seem necessary. In practice, however, police and other law-enforcement entities have great leeway when it comes to adding people to the database; as a result, teachers, coaches, and relatives of suspected gang members could themselves be listed. Kamara pointed out that investigations found 42 infants under the age of one listed in CalGang, as well as other minors whose families were not informed of their inclusion, as required by law. Once they are added to the database, individuals have had trouble getting themselves removed — or even learning that they are on the list [3].

Due to these issues, the Los Angeles Police Department announced in 2020 that it would no longer use CalGang [2]. However, other California police departments still utilize this application, and Immigration and Customs Enforcement (ICE) employ the similarly problematic ICEGangs database as part of the deportation decision process.

Know Your Enemy

Part of the difficulty in combating such technological problems is that they often work as designed; their intrinsically unequal effects are part of the package. "If you're an American living in the U.S., drones are fantastic," Kamara said, referencing package delivery and recreational use. "But if you live in a different part of the world, like in Pakistan, you have a very different feeling about drones. They affect you very differently [because] they're weapons of war."

This disparity also applies to other types of robots, which are heavily funded by the military or law enforcement. Even though these machines—such as the "dancing dog" robots of viral video fame—may have life-saving and life-enhancing applications, they are already being weaponized. One's opinion of such robots largely depends on whether the technology is helping or harming them.

To better explain the situation, Kamara used a metaphor from his own specialization of cryptography: adversarial models, wherein someone attempts to learn the contents of an encrypted message. The adversary in this analogy is a person or group who wants to discover other people's secrets or harm them in some way.

"When you're designing a system, you have multiple kinds of adversaries with different powers and different goals," Kamara said. "If I'm young and Black and live in St. Louis or New York City, odds are that the police are part of my adversarial model. When I leave my house, I have to think about my interactions with police, how to survive those interactions, or whether I'm going to be harassed."

This is not the case for many white people, whose primary interactions with police are neutral or positive. They thus have different adversarial models than Black people. Similarly, women have different adversarial models than men, immigrants—particularly immigrants of color—have different adversarial models than citizens, and so forth.

"Technology is produced with certain adversarial models in mind," Kamara said. "But all of these other groups and communities just don't come up, so the problems that they face are not being met by the technology being produced. This is definitely the case with respect to privacy, security, and safety."

Why Stand on a Silent Platform?

Some programmers may be loath to accept that one does not have to be ideologically racist to produce something that can be used in racist ways. Both Kamara and Venkatasubramanian emphasized that combatting a racist and exploitative system requires *active* opposition, rather than ideological neutrality.

"In computer science, we don't think of ourselves as being part of a larger system of societal governance," Venkatasubramanian said. "We think of ourselves as tool builders. One of the difficulties has been to realize that 90 percent of the hammers we're building are being used to beat on people, and only 10 percent are being used to beat on nails."

Acknowledging that neutrality is not a binary state is one step toward understanding unintended consequences. "It's not helpful to talk about the artifact itself as being neutral or not," Venkatasubramanian continued. "It's an end stage of a whole process that involves people who have their own judgments about what they should be doing."

For obvious reasons, technologists like to create technological solutions for every issue, regardless of appropriateness. While Kamara admitted to such bias himself, he also acknowledged that a more diverse field would mitigate partiality. "A huge part of the solution is that we just need more diversity," he said. "We need people with different life experiences. They know what the problems are and can come up with good solutions."

However, hiring more white women and people of color to produce code will not solve these issues if the field's established professionals do not admit that problems exist. To that end, a broader education that encompasses history, sociology, and policy would greatly benefit technologists. Furthermore, software that is harmful towards other people might negatively affect white male computer scientists in the future [1].

Problematic technologies like CalGang and surveillance drones are unlikely to vanish because ample funding for their creation continues to flow. However, understanding their potential to hurt certain groups of people can shift value systems in computer science, much like how the natural sciences grapple with funding sources and research independence. "Not only do we need to participate [in change] because of what we have wrought," Venkatasubramanian said. "We should participate because we could do so much more."

References

[1] Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Medford, MA: Polity.

[2] Chabria, A., & Miller, L. (2020, June 24). Reformers want California police to stop using a gang database seen as racially biased. *Los Angeles Times*. Retrieved from https://www.latimes.com/california/story/2020-06-24/california-police-urged-to-stop-using-gang-database-deemed-biased.

[3] Winston, A. (2016, March 23). You may be in California's gang database and not even know it. *Reveal*. Retrieved from https://revealnews.org/article/you-may-be-in-californias-gang-database-and-not-even-know-it.

Matthew R. Francis is a physicist, science writer, public speaker, educator, and frequent wearer of jaunty hats. His website is BowlerHatScience.org.

---

[1] https://oag.ca.gov/calgang

[2] https://www.aaas.org/events/2021-aaas-annual-meeting

Renewables Reliability

Continued from page 2

enable novel solutions through customer cooperation. An active research strand is tackling these problems with mean-field stochastic models and principal-agent models that were recently developed by economists and financial mathematicians. A paradigmatic example of modeling distributed local storage via the mean-field game (MFG) framework published last year [3]. The authors of the study employed the MFG approximation to design a price-based signal that aligns consumer behavior with the interests of grid balancing.

Similarly, researchers utilized recent developments in continuous-time optimal contract theory to create an implementable demand response contract that allows the utility to control the responsiveness of a pool of consumers, thus significantly increasing demand response's reliability when necessary [2]. The potential of demand response was evidenced in California in August 2020, when vigorous social media campaigns for short-term electricity conservation helped the state avoid rolling brownouts that were trigged by record heat waves. When activated ad hoc in an emergency setting, the use of internet-based, decentralized tools to shape electricity demand has the potential to reduce the need for costly batteries or the construction of new fossil fuel plants. A related topic that is highly amenable to optimization and stochastic techniques is the design and implementation of smart charging networks that would adaptively charge electric vehicles overnight to maximally benefit overall grid management.

References

[1] Advanced Research Projects Agency – Energy. *Performance-based energy resource feedback, optimization, and risk management (PERFORM)*. Retrieved from https://arpa-e.energy.gov/technologies/programs/perform.

[2] Aïd, R., Possamaï, D., & Touzi, N. (2017, June 1). Electricity demand response and optimal contract theory. *SIAM News*, *50*(5), p. 1.

[3] Alasseur, C., Tahar, B., & Matoussi, A. (2020). An extended mean field game for storages in smart grids. *J. Optim. Theory Appl., 184*, 644-670.

[4] Bloomberg News. (2020, April 25). *The 20 minutes that broke the U.S. oil market*. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2020-04-25/the-20-minutes-that-broke-the-u-s-oil-market.

[5] California Independent System Operator. *Managing oversupply*. Retrieved from http://www.caiso.com/informed/Pages/ManagingOversupply.aspx.

[6] Cicala, S. (2020). Powering work from home. *NBER working paper series*. National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w27937.

[7] Coulon, M., Powell, W., & Sircar, R. (2013). A model for hedging load and price risk in the Texas electricity market. *Energy Econ., 40*, 976-988.

René Aïd is a professor of economics in the Department of Economics at Paris Dauphine University. Mike Ludkovski is a professor in the Department of Statistics and Applied Probability at the University of California, Santa Barbara. Ronnie Sircar is a professor in the Department of Operations Research and Financial Engineering at Princeton University.

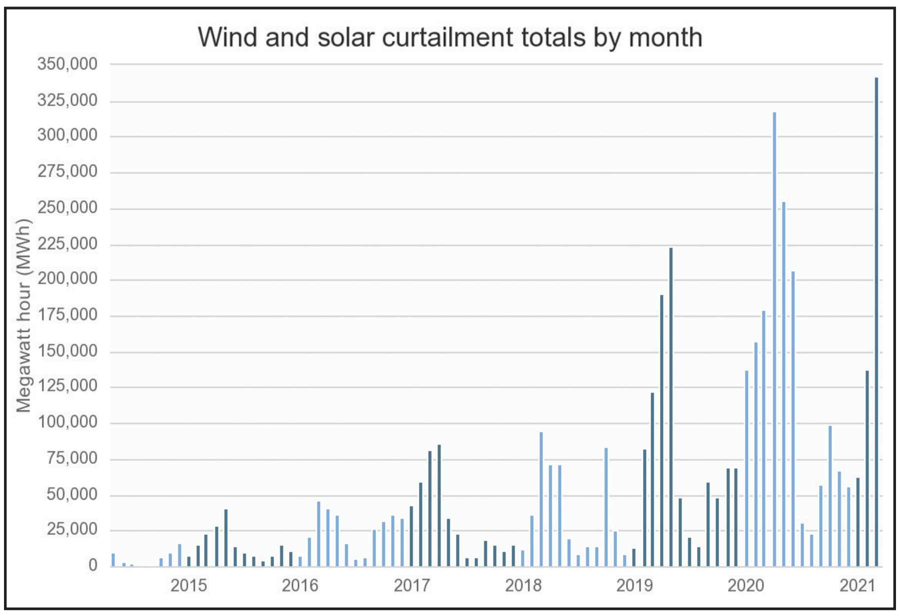


Figure 2. *California Independent System Operator (ISO) monthly curtailment of renewable generators. Curtailment records were established in 2020. Figure courtesy of [5] and licensed with permission from the California ISO. Any statements, conclusions, summaries or other commentaries expressed herein do not reflect the opinions or endorsement of the California ISO.*

The Exact Solution of All of Continuum Mechanics

By Richard D. James, Alessia Nota, Gunjan Pahlani, and Juan J. L. Velázquez

Let A be any 3×3 matrix, assume t is small enough that $\det(I+tA)>0$, and consider the function

$$v(x,t)=A(I+tA)^{-1}x, \quad x\in\mathbb{R}^3. \tag{1}$$

Make one of the many choices of A such that $\operatorname{div} v=0$ and substitute into the Navier-Stokes equations of fluid mechanics:

$$\frac{\partial v}{\partial t}+(v\cdot\nabla)v=-\nabla p+\nu\Delta v. \tag{2}$$

Choosing for example $p=p(t)$, a quick calculation shows that $v(x,t)$ is an exact solution. In fact, both sides of the Navier-Stokes equations vanish. Equation (2) is for the incompressible case, but $v(x,t)$ also solves the Navier-Stokes equations for a compressible fluid; for the latter, choose any A and the corresponding density given by $\rho(x,t)=\rho_0\det(I+tA)^{-1}$, $\rho_0=const.>0$.

Now try the equations of elasticity (linear or nonlinear). Again, they yield an exact solution for every A. To witness this phenomenon in the general nonlinear case, we must first convert from the Eulerian to the Lagrangian description of motion by solving the system of ordinary differential equations (ODEs) $\dot{y}(z,t)=v(y(z,t),t)$, $y(z,0)=z$; the solution is $y(z,t)=(I+tA)z$. We then immediately see that the following holds for every elastic material:

$$\rho_0\frac{\partial^2 y}{\partial t^2}=\operatorname{div}T, \quad T=\tilde{T}(\nabla y). \tag{3}$$

At first glance, this is quite surprising. Aren't solutions of partial differential equations supposed to depend on the choice of coefficients? Why should a particular solution for air also work for water, rubber, and steel? Very different physics—and very different kinds of atomic forces—underly the flow of a gas or liquid when compared to the forces that bind a crystalline solid. Why is the exact same nine-parameter family of functions possible in both scenarios? The motions that $v(x,t)$ describe are not trivial; they may be volume preserving, they can have vorticity that changes in time, they have a strong singularity when $\det(I+tA)\to 0$, and they can most certainly exist far from equilibrium.

How about something more exotic? Consider any one of the accepted but highly nonlinear models of non-Newtonian fluids, the theory of liquid crystals, or nonlinear viscoelasticity. Or think about one of the models of phase transformations with a change of type (regularized or not). These examples again produce an exact solution for every A.

What about thermodynamics? Here we finally find a departure from universality. In all of the aforementioned cases—augmented by the first and second laws of thermodynamics—$v(x,t)$ still satisfies the laws of conservation of mass and momentum, with temperature $\theta(t)$ assumed to be a function of time only. But the energy equation becomes an ODE for $\theta(t)$. The coefficients of this ODE depend on the material, and the corresponding evolution of temperature is material dependent.

Symmetry often accounts for coincidences like this, as is the case here. So, where is the symmetry group?

It is best to descend to atomistic theory in order to answer this question. Let us consider the molecular dynamics (MD) of atoms with positions $y_k(t)$ and masses m_k. These satisfy the equations of MD:

$$m_k\ddot{y}_k(t)= -\frac{\partial\varphi}{\partial y_k}(\dots,y_{i-1}(t),y_i(t),y_{i+1}(t),\dots). \tag{4}$$

The well-known symmetries of quantum mechanics—the frame indifference and permutation invariance of atomic forces—are here. We wish to examine cases wherein infinitely many atoms fill space, analogous to the function $v(x,t)$, $x\in\mathbb{R}^3$. The potential energy φ is typically infinite in these cases, but the infinite system of ODEs of MD can still make perfect sense. We would therefore like to express these symmetries in terms of the forces $-\partial\varphi/\partial y_k$. The fundamental symmetries are

$$\frac{\partial\varphi}{\partial y_k}(\dots,Qy_{i-1}+c,Qy_i+c,Qy_{i+1}+c,\dots) = Q\frac{\partial\varphi}{\partial y_k}(\dots,y_{i-1},y_i,y_{i+1},\dots)$$

(frame indifference)

$$\frac{\partial\varphi}{\partial y_k}(\dots,y_{\Pi(i-1)}(t),y_{\Pi(i)}(t),y_{\Pi(i+1)}(t),\dots) = \frac{\partial\varphi}{\partial y_{\Pi(k)}}(\dots,y_{i-1},y_i,y_{i+1},\dots)$$

(permutation invariance),

where $Q\in\mathrm{O}(3)$, $c\in\mathbb{R}^3$, and $\Pi(\cdot)$ is a permutation that preserves the species of atom.

Both O(3) and the translation group are continuous groups, but we have discrete atomic positions. This fact suggests that we should examine discrete groups of isometries, i.e., elements $(Q\,|\,c)$—often written in this notation—with $Q\in\mathrm{O}(3)$, $c\in\mathbb{R}^3$. These can generate many groups of the form $\mathcal{G}=\{(Q_1\,|\,c_1),(Q_2\,|\,c_2),\dots,(Q_N\,|\,c_N)\}$, some of which are listed in the International Tables for Crystallography.[1] N can be infinite, and the group product is $(Q_1\,|\,c_1)(Q_2\,|\,c_2)=(Q_1Q_2\,|\,c_1+Q_1c_2)$. One last crucial assumption is that we will allow $c_1,c_2,\dots$ to depend on time, but we only allow affine time dependence: $c_1=a_1t+b_1$, $c_2=a_2t+b_2$, $c_3=a_3t+b_3,\dots$. The a_i and b_i must be chosen so that $\mathcal{G}$ remains a group for, say, all $t>0$. Although this vaguely resembles a Galilean transformation, it is not because the a_i, b_i vary between group elements.

Consider any (discrete) isometry group $\mathcal{G}$—possibly time dependent as described—and designate a subset of atoms with positions $y_1(t),\dots y_M(t)$ as the *simulated atoms*, with initial positions $y_1^\circ,\dots,y_M^\circ$ and initial velocities $v_1^\circ,\dots,v_M^\circ$. We obtain the other atom positions by applying $\mathcal{G}$ to the simulated atoms using the obvious rule: $Q_1y_1(t)+c_1,Q_1y_2(t)+c_1,\dots,Q_2y_1(t)+c_2$, $Q_2y_2(t)+c_2,\dots,Q_Ny_1(t)+c_N,Q_Ny_2(t)+c_N,\dots$. Now we write the MD equations from above, but only for M simulated atoms. On the right-hand side of (4), substitute these formulas for all other *nonsimulated atoms*. Because the nonsimulated atoms are given by formulas in terms of the simulated atoms, the equations for the simulated atoms become a system of (nonatonomous) ODEs in standard form. The necessary initial conditions are $y_1^\circ,\dots,y_M^\circ$, $v_1^\circ,\dots,v_M^\circ$. We then solve these equations. A straightforward theorem reveals that even though the nonsimulated atoms are given by formulas, these atoms exactly satisfy the MD equations for their forces; we utilize both frame indifference and permutation

See **Continuum Mechanics** *on page 7*

---

[1] https://it.iucr.org

Continuum Mechanics

Continued from page 6

invariance to show this. Figure 1 depicts several snapshots of a simulation.

Now we are ready to explain $v(x,t)$. We select the simplest time-dependent translation group, with elements that can be written as $(I|(I+tA)(\nu^1 e_1+\nu^2 e_2+\nu^3 e_3))$. Here, ν^1,ν^2,ν^3 are integers and e_1,e_2,e_3 are linearly independent. This is clearly a group, and it has the affine time dependence. We can also determine the macroscopic velocity field of this solution with minimal difficulty: it is exactly $v(x,t)$. A clear victory for continuum mechanics!

By examining this argument more closely, we can see that it really pertains to a largely unstudied, explicit, and time-dependent invariant manifold of the MD equations. In fact, lots of groups, time dependencies, and these types of invariant manifolds are present. Their forms are independent of the material; the same manifolds apply to air, water, and steel.

We now arrive at the article's main purpose. Recall that the molecular density function $f(t,x,v)$ of the kinetic theory of gases represents the probability density of finding an atom with velocity v in a small neighborhood of x at time t, in Eulerian form. Let us stay with the time-dependent translation group and examine the statistics of the MD solutions. Draw a ball $\mathcal{B}_0$ of any diameter that is centered at the origin. Now choose integers ν^1,ν^2,ν^3 and draw a ball $\mathcal{B}_\nu$ of the same diameter that is centered at $x=(I+tA)\,(\nu^1 e_1+\nu^2 e_2+\nu^3 e_3)$. Since the simulated atoms quickly diffuse into the nonsimulated atoms during a simulation (see Figure 1), it is not unusual for $\mathcal{B}_0$ and $\mathcal{B}_\nu$ to contain both simulated and nonsimulated atoms at any particular time. The velocities of atoms in $\mathcal{B}_0$ and $\mathcal{B}_\nu$ are different. Nevertheless, if we know the velocities of atoms in $\mathcal{B}_0$, we can then use explicit formulas to calculate the velocities of atoms in $\mathcal{B}_\nu$. But based on its interpretation, this calculation implies an ansatz for the molecular density function f: $f(t,x,v)=g(t,v-A(I+tA)^{-1}x)$. Ignore the fact that x was a special point and substitute the ansatz into the general form of the Boltzmann equation. It yields an exact reduction.

Notice that we lose the periodicity regardless of whether we pass from MD to the reduced Boltzmann equation or to continuum mechanics. This is comforting.

We have been studying this reduced equation for g and simulating the invariant manifold of MD itself. We do not believe that anything like a statistical description of nonequilibrium exists in general. After all, the Boltzmann equation—which treats only the simplest kind of material—gives an initial value problem for a nonlinear integro-differential equation in seven variables. There is no chance that a simple reduction to an explicit form of f exists far from equilibrium. However, *on the manifold given by* g*—or perhaps on one of the invariant manifolds of the aforementioned general equations of MD—a kind of explicit non-equilibrium statistical mechanics might exist* that approaches the simplicity of equilibrium statistical mechanics, which also lies on an invariant manifold $(H=const)$. Finding this, even in special cases, would be a tremendous achievement. Our work on the Boltzmann equation perhaps gives some hints, but more ideas are necessary.

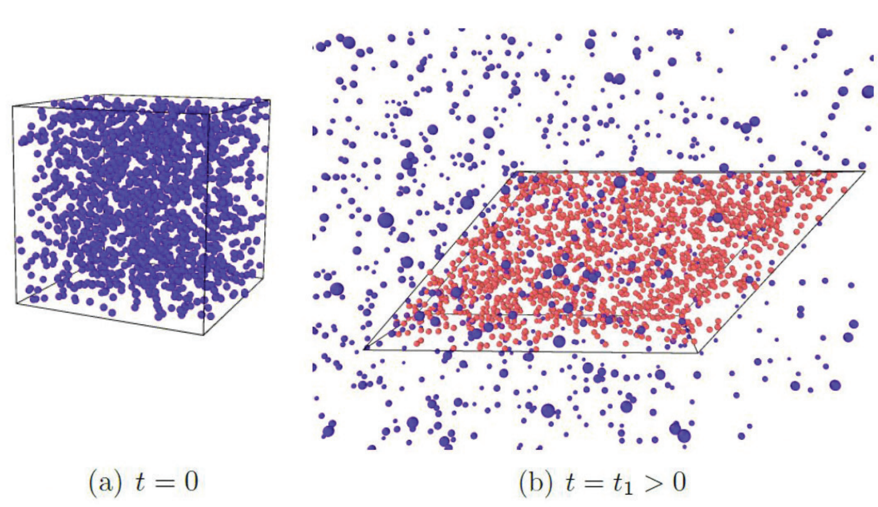


Figure 1. *Two snapshots from a simulation that employs Lennard-Jones argon and* $A=\kappa(e_1\otimes e_1+e_1\otimes e_2)$ *with* e_1,e_2 *orthonormal, ultimately yielding a macroscopic motion with extension and shear.* **1a.** *The initial state depicts the simulated atoms (blue).* **1b.** *The nonsimulated atoms (red) fill the entire space, but for clarity only those in the parallelepiped are shown; this parallelepiped is deformed from the cube according to the macroscopic motion. The larger simulated atoms are closer to the reader. Note that the simulated atoms quickly diffuse into the sea of nonsimulated atoms. A full animation is available online. Figure courtesy of Gunjan Pahlani.*

Further Reading

[1] Bobylev, A., Nota, A., & Velázquez, J.J.L. (2020). Self-similar asymptotics for a modified Maxwell–Boltzmann equation in systems subject to deformations. *Comm. Math. Phys.*, *380*(1), 409-448.

[2] Dayal, K., & James, R.D. (2010). Nonequilibrium molecular dynamics for bulk materials and nanostructures. *J. Mech. Phys. Solids*, *58*(2), 145-163.

[3] James, R.D., Nota, A., & Velázquez, J.J.L. (2019). Long-time asymptotics for homoenergetic solutions of the Boltzmann equation: Collision-dominated case. *J. Nonlin. Sci.*, *29*(5), 1943-1973.

[4] James, R.D., Nota, A., & Velázquez, J.J.L. (2019). Self-similar profiles for homoenergetic solutions of the Boltzmann equation: Particle velocity distribution and entropy. *Arch. Ration. Mech. Anal.*, *231*(2), 787-843.

[5] James, R.D., Nota, A., & Velázquez, J.J.L. (2020). Long time asymptotics for homoenergetic solutions of the Boltzmann equation. Hyperbolic-dominated case. *Nonlin.*, *33*(8), 3781.

[6] Pahlani, G., Torres, E., Schwartzentruber, T.E., & James, R.D. (2021). Objective molecular dynamics of dissociating nitrogen under high temperature conditions. In *AIAA Scitech 2021 Forum* (p. 0707).

Richard D. James teaches continuum mechanics at the University of Minnesota. Comments can be addressed to him at james@umn.edu. Alessia Nota is a tenure-track assistant professor at the University of L'Aquila (DISIM). She was previously a postdoctoral researcher at the University of Bonn and the University of Helsinki. Her research area is non-equilibrium statistical mechanics, with a particular focus on the kinetic theory of gases and plasmas. Gunjan Pahlani is a Ph.D. student in the Department of Aerospace Engineering and Mechanics at the University of Minnesota. Her research focuses on the formulation of computational methods at the molecular level that exploit long-range symmetries for the simulation of hypersonic gas flows and plastic deformation in materials. Juan J. L. Velázquez is a professor in the Institute for Applied Mathematics at the University of Bonn, prior to which he was a professor in the Department of Applied Mathematics at the Complutense University of Madrid and a postdoctoral Fulbright Fellow at the University of Minnesota. His research focuses on several problems of partial differential equations, particularly in the study of kinetic equations.

CSE21 Panel Explores the Importance of Mentorship

By Jillian Kunze

Mentors can help students and early-career mathematicians navigate the complex world of research, build strong professional relationships, balance multiple responsibilities, and plan for the future. During the 2021 SIAM Conference on Computational Science and Engineering[1] (CSE21), which took place virtually this March, the Sustainable Horizons Institute's Broader Engagement program[2] sponsored a mentoring panel[3] and networking session to explore this idea. Christine Harvey (MITRE Corporation) moderated the discussion between two mentor-mentee pairs: Sally R. Ellingson (University of Kentucky) and Derek Jones (Lawrence Livermore National Laboratory), and Jay Lofstead (Sandia National Laboratories) and Paula Fernanda Olaya Garcia (University of Tennessee, Knoxville). The panelists discussed the importance of mentorship in applied mathematics and presented strategies for maintaining such relationships.

Ellingson utilizes her experience with high-performance computing to perform cancer research at her university; she also routinely visits Lawrence Livermore and Lawrence Berkeley National Laboratory. She met Jones—who received both his bachelor's and master's degrees from the University of Kentucky—while interviewing students for a research assistantship in 2016. Now, Jones spends his time developing deep learning for cancer research at Lawrence Livermore and pursuing his Ph.D. at the University of California, San Diego.

Jones' involvement with Ellingson's work served as an exciting introduction to the world of biology and high-performance computing. "Her project exposed me to a whole new world of research," he said. "She has opened a lot of doors for me, and I have tried to take each opportunity and make the best of it." Jones had no research experience when he began working with Ellingson, so she helped him learn to navigate the research process. In turn, Ellingson noted that collaborating with Jones has taught her much about what it means to be a mentor.

Lofstead, who researches data management for supercomputers at Sandia, met mentee Olaya Garcia through her university advisor. Olaya Garcia received her bachelor's degree from the Pontificia Universidad Javeriana in Bogotá, Colombia and is now a graduate student in computer science at the University of Tennessee, Knoxville. Among other things, she focuses on enhancing the reproducibility of scientific workflows with cutting-edge container technology.

Lofstead believes that it is important to really get to know one's mentees by discussing both research interests and personal endeavors. He begins the process by getting a sense of a mentee's capabilities, then pushes them to succeed while offering continual support. "I never ask anyone to do anything that I don't believe they will be successful at, even if they have no idea how to do it," Lofstead said. Both he and Olaya Garcia agreed that the mentee should be primarily responsible for developing and maintaining an action plan for research; the mentor can help, especially at the beginning, but the student should gradually take more control. "Lofstead gives a lot of responsibilities, but he also gives a lot of trust and authority," Olaya Garcia said. In addition, both parties should be realistic with their goals and communicate their expectations of one another.

Harvey then opened the session to attendee questions and asked how to best find a mentor outside of a specific program's framework. When Lofstead looked for mentors in the past, he reached out to individuals at a career level that he hoped to attain in 10 years. "If you see someone doing the things you want to do—even if it is purely aspirational—just approach them and say, 'I want to do what you do, can you help me get there?'" he said. Furthermore, a potential mentor does not necessarily have to do exactly what a mentee wishes to achieve — one might simply choose somebody that they admire and with whom they work well. "One of the biggest things is having that sense of trust and understanding of each other," Jones said.

When searching for a mentor or mentee, one should account for several important considerations. Mentees must be willing to trust their mentors and remain passionate about their work. Olaya Garcia noted that mentors who offer guidance and support in multiple areas are especially valuable. "It is best to work with someone who is not only focused on work, but will understand the struggles of being in a Ph.D. program," she said. Having a mentor who appreciates work-life balance and recognizes the necessity of periodic breaks is indispensable. Olaya Garcia also applies this mindset herself when she works with undergraduates; she draws from her own experiences at that stage and recognizes that students still have a lot to learn.

From left to right: Jonathan Allen (Lawrence Livermore National Laboratory), Julie Mitchell (Oak Ridge National Laboratory), Sally Ellingson (University of Kentucky), and Derek Jones (Lawrence Livermore National Laboratory) pose after presenting at a minisymposium about computational drug discovery during the 2019 SIAM Conference on Computational Science and Engineering, which took place in Spokane, Wash. Ellingson, who serves as a mentor to Jones, met him when he was interviewing for a research assistantship in 2016. They both spoke about the value of mentorship at the virtual 2021 SIAM Conference on Computational Science and Engineering (CSE21) earlier this year. Photo courtesy of Mary Ann Leung.

In today's climate, many connections are formed online. However, virtually reaching out to a potential mentor can be intimidating. Ellingson recommended including a personal message when connecting via email or through LinkedIn. "You need to put effort into it to make sure that you get something out of it," she said. One's initial greeting should not seem like a blast message that was sent to a lot of people; it should contain at least a few sentences that are specific to the recipient. The message should be personal and straightforward, and preferably readable in under two minutes. If email is insufficient, Lofstead recommended asking whether any existing connections can facilitate an introduction to the person in question.

The popularity of interdisciplinary majors is on the rise, and mentors can help students shape their own individual paths. The panelists advised students to spend time exploring different research directions. "Give yourself a wide variety of things to work on and try, and figure out what you like," Ellingson said. Though a career path is a uniquely personal decision, conversing with experienced professionals who have explored dissimilar career trajectories can help students learn about various fields and identify possible future directions.

Furthermore, mentors should acknowledge and respect the questions and priorities of their mentees, especially those from diverse backgrounds and underrepresented communities. Lofstead, who was a first-generation college student, affirmed that addressing questions and issues outside of technical research is essential for successful relationships. Jones echoed this sentiment. "A nontrivial part of the difficulty of graduate school is these challenges, which minorities might experience more than other population groups," he said. Because issues in graduate school and early-career settings affect different people in different ways, junior mathematicians should seek out mentors who can relate to these unique experiences. "There is a context behind every person," Olaya Garcia said. "You need to see an individual not only as the one that is doing the research, but as a whole person."

Mentor-mentee pairs must also establish the frequency of their communications, especially when they are unable to connect in person. Some students may require an increased level of interaction, or less interaction and more space to conduct their work. Ellingson always follows her students' leads when it comes to productivity strategies. Though sending an email may be easiest when working from home, arranging a dedicated time to meet can enable a more robust back-and-forth discussion. For example, Jones has been making a concerted effort to schedule times to converse, as keeping up with colleagues remotely is more difficult than in person.

After the question-and-answer period with the panelists, the session moved into gather.town[4] for networking purposes. Attendees entered a virtual mentor mixing room, where they chatted with the panelists to learn even more about building and maintaining mentoring relationships.

CSE21 attendees can access a recording of this panel and other conference presentations by logging in to the virtual platform. Recordings are available until June 5th. If you weren't able to attend the meeting, SIAM has reopened registration at a discounted rate for on-demand viewing of all talks. Learn more and register online.[5]

Jillian Kunze is the associate editor of SIAM News.

---

[1] https://www.siam.org/conferences/cm/conference/cse21

[2] http://shinstitute.org/siam-cse21-broader-engagement-program

[3] https://meetings.siam.org/sess/dsp_talk.cfm?p=107475

[4] https://gather.town

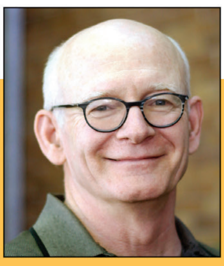
[5] https://www.siam.org/conferences/cm/registration/cse21-registration

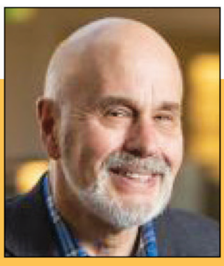
SIAM Fellows

2021

SIAM is pleased to announce the newly selected Class of SIAM Fellows—a group of distinguished members of SIAM who were nominated by their peers for exceptional contributions to the fields of applied mathematics and computational science. Please join us in congratulating these 28 members of our community.


Alejandro Aceves
Southern Methodist University


James Burke
University of Washington


Robert Calderbank
Duke University


Xiaojun Chen
Hong Kong Polytechnic University

siam Fellow

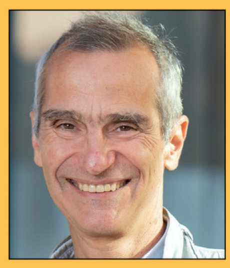

Edmond Chow
Georgia Institute of Technology

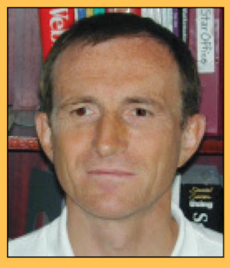

Robert Falgout
Lawrence Livermore National Laboratory


Martín Farach-Colton
Rutgers University


Shmuel Friedland
University of Illinois at Chicago


Gary Froyland
University of New South Wales


Tryphon Georgiou
University of California, Irvine


Jean-Luc Guermond
Texas A&M University


Trachette Jackson
University of Michigan


Jeremy Kepner
MIT Lincoln Laboratory

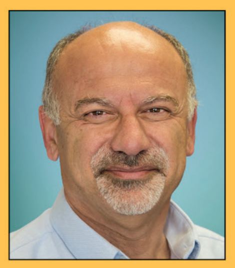

Denise Kirschner
University of Michigan


Rachel Levy
American Mathematical Society / AAAS


Per-Gunnar Martinsson
University of Texas at Austin


Anna Mazzucato
Penn State University


Kirsten Morris
University of Waterloo


Habib Najm
Sandia National Laboratories


Qing Nie
University of California, Irvine


Beatrice Riviere
Rice University


Jonathan Rubin
University of Pittsburgh

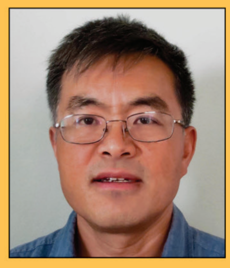

Jennifer Scott
University of Reading and Science and Technology Facilities Council


Eitan Tadmor
University of Maryland, College Park


Shang-Hua Teng
University of Southern California


Andreas Wächter
Northwestern University


Rebecca Willett
University of Chicago


Jack Xin
University of California, Irvine

Learn more about the 2021 Class of SIAM Fellows and nominate a colleague by October 20, 2021 for the 2022 Class of Fellows:
www.siam.org/Prizes-Recognition/Fellows-Program

siam. Society for Industrial and Applied Mathematics

Hanging Cables and Hydrostatics

MATHEMATICAL CURIOSITIES

By Mark Levi

Tension Is Linear

When driving down a country road and observing hanging electric cables by the roadside, I marvel at nature's ability to solve a minimization problem; out of all possible shapes, it finds the one with the least potential energy. These hyperbolic cosine-shaped cables have another interesting property: their tension depends linearly on the height h (see Figure 1):

$$T - T_0 = \rho g h. \tag{1}$$

Here, ρ is the cable's linear density — i.e., the mass per unit length. This is reminiscent of Pascal's law, $p - p_0 = \rho g h$, for the water pressure at depth h; in this case—unlike in (1)—ρ stands for the water's density. It is not a coincidence; one can think of the hanging rope as a one-dimensional fluid wherein the tension corresponds to pressure in the water and unstretchability corresponds to incompressibility. In fact, the same energy-conservation argument proves both (1) and Pascal's law. The argument goes as follows. I begin by holding a stationary chain by its two ends, then advance each end by small distance ds in such a way that every particle of the chain advances along the curve (see Figure 2).

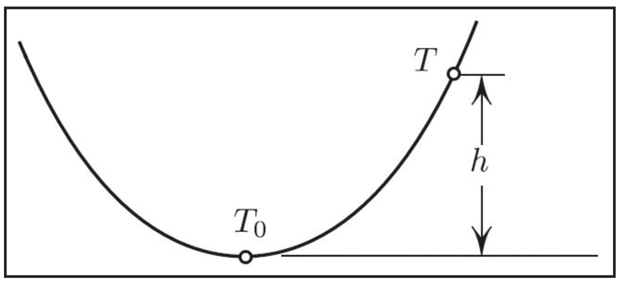


Figure 1. *Tension in a hanging cable varies linearly with height.*

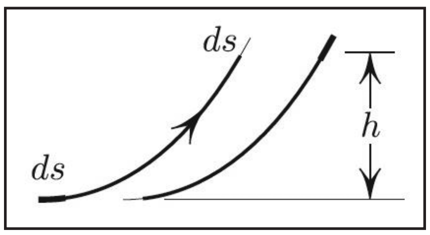


Figure 2. *Advancing the cable by ds changes its potential energy by $dm\,gh = \rho ds\,gh$ and takes work $(T - T_0)ds$.*

The advancing hand did work $T\,ds$ and the retreating hand did work $-T_0\,ds$; the minus is due to the fact that the applied force points against the displacement. The result of this motion is now the same as simply raising the element ds by height h, with the change of potential energy $dm\,gh = \rho ds\,gh$. Therefore,

$$T\,ds - T_0\,ds = \rho ds\,gh,$$

which amounts to (1). One can apply the exact same argument to prove Pascal's law, although textbooks do not usually take this approach.

Curvature and Tension

Can one "see" the tension T_0 at the lowest point of the cable? If one knows ρ (the linear density), the answer is yes. T_0 is the radius R of curvature, up to a factor:

$$T_0 = \rho g R. \tag{2}$$

Indeed, the weight of the segment ds in Figure 3 is supported by the vertical tension:

$$T \sin d\theta = \rho g ds \quad \text{or} \quad T\frac{\sin d\theta}{ds} = \rho g.$$

With $ds \to 0$, this becomes $T_0 k = \rho g$ — where k is the curvature. This amounts to (2).

By combining (1) and (2), we get $T = m\rho g(R + h)$; that is, the tension equals the weight of the cable of length $R + h$.

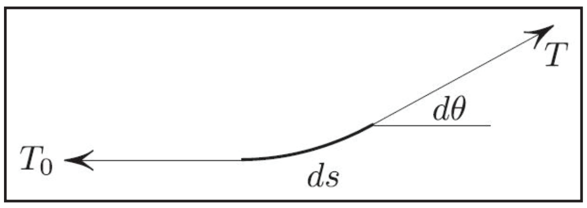


Figure 3. *Proving that tension at the lowest point is proportional to the radius of curvature.*

Area and Length

The catenary—i.e., the graph of the hyperbolic cosine—has a remarkable property: the area under the arc above any interval equals its length (see Figure 4). That is,

$$\int_0^x f(t)dt = \int_0^x \sqrt{1 + (f'(t))^2}\,dt \tag{3}$$

if $f = \cosh$.

One can either check this by substitution or deduce it by differentiating (3) and solving the resulting ordinary differential equation (ODE) for $y = f(x)$:

$$y' = \sqrt{y^2 - 1}, \quad y(0) = 1. \tag{4}$$

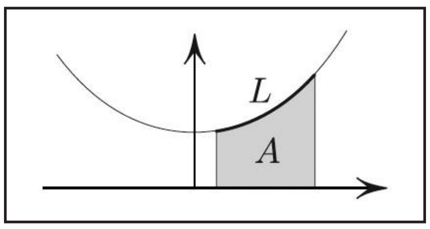


Figure 4. *For the catenary, $A = L$.*

The initial condition $y(0) = 1$ results from substituting $x = 0$ into the derivative of (3). Separation of variables and some manipulation leads to $y = \cosh x$.

Uniqueness

Is cosh the only function with this property? A quick reflection—or a look at (3)—shows that the constant $f(x) \equiv 1$ has the same property. This initially worried me; where did I lose the answer when solving (4)? Having two solutions for the same ODE signals that the uniqueness theorem does not apply; indeed, $\sqrt{y^2 - 1}$ fails assumptions of every uniqueness theorem at $y = 1$. With the two solutions $f(x) = \cosh x$ and $f(x) \equiv 1$, infinitely many others must also exist (according to a theorem of Kneser and Zaremba). However, these solutions are not very interesting; they are simply concatenations that are defined, for an arbitrary c, by $f(x) \equiv 1$ for $x \in [0, c]$ and $f(x) = \cosh(x - c)$ for $x > c$.

The figures in this article were provided by the author.

Mark Levi (levi@math.psu.edu) is a professor of mathematics at the Pennsylvania State University.

CSE21 Session Addresses Early Careers in Academia, National Laboratories, and Industry

By Jillian Kunze

During the 2021 SIAM Conference on Computational Science and Engineering[1] (CSE21), which took place virtually this March, an early career panel[2] explored several possible occupational directions for recent graduates in applied mathematics. Kevin Carlberg (Facebook) and Victoria Howle (Texas Tech University) led the discussion, which addressed the advantages and challenges of various career paths. Panelists spoke in three distinct groups: academia, national laboratories, and industry. Attendees—the majority of whom were graduate students—had the opportunity to converse with the participants and glean their advice on early careers in these three areas.

Careers in Academia

Malena I. Espanol (Arizona State University), Stefan J. Kollet (Agrosphere Institute), and Miriam Mehl (University of Stuttgart) shared their respective paths through academia, beginning with impactful decisions that they made early in their careers. Mehl was glad to have joined a research consortium that helped her learn about different fields and meet her peers. "Networking should not be underestimated," she said. Espanol also emphasized the importance of broad experiences. "If you get the chance, take very diverse classes," she advised. Summer internships in a different area of study—possibly in industry or at a national laboratory—can also widen one's horizons. In fact, Kollet chose to travel to a new country for his Ph.D. and postdoctoral research, which exposed him to an entirely different system of academia.

Panelists urged attendees who were about to enter graduate school to seek out advisors who share their priorities. To do so, students should investigate a prospective advisor's productivity, look into the achievements of their former students, and chat with group members. "There are no good or bad environments, with a few exceptions" Mehl said. "It just has to be a fit between the students and the supervisors."

There are both advantages and disadvantages to postdoctoral research positions versus tenure-track posts after graduate school. Postdoctoral positions provide relevant experience but do not work with everyone's personal situations. However, the many duties of a tenure-track position can constitute an abrupt change from being a student. "A faculty position is more than just research," Espanol said. "It consists of teaching, supervising students, and applying for grants; it is a lot of work that takes away from research time." When interviewing for a tenure-track job in academia, one should negotiate for additional considerations beyond salary, such as money for students and postdocs, a publishing budget, and multiple rooms. "This kind of negotiation is really important, as it is one of the few chances you will get to do so," Kollet said. "You should take this chance and take it very seriously."

CAREERS IN
MATHEMATICAL
SCIENCES

Postdoctoral researchers often wonder whether they should focus on projects that are closely related to their Ph.D., or extend their skills and expertise. Because applied mathematicians frequently work with a wide variety of applications, it is generally beneficial to extend one's capabilities. "You need to have good knowledge of a few fields so that you can talk to other people and supervise Ph.D. students who want to go in different directions," Kollet said. "In academia, you have to broaden your perspective." To accomplish this, postdocs can look for fields in which knowledge and methods from their Ph.D. are still applicable.

A panel at the 2021 SIAM Conference on Computational Science and Engineering (CSE21), which took place virtually in March, addressed early career paths for mathematicians in academia, national laboratories, and industry. Top row, left to right: Malena Espanol (Arizona State University), Stefan Kollet (Agrosphere Institute), and Miriam Mehl (University of Stuttgart). Middle row, left to right: Jacob Schroder (University of New Mexico), Heidi Thornquist (Sandia National Laboratories), and Juliane Mueller (Lawrence Berkeley National Laboratory). Bottom row, left to right: Julia Ling (Citrine Informatics), Maxim Naumov (Facebook), and Jeff Hammond (NVIDIA).

Careers in National Laboratories

Jacob B. Schroder (University of New Mexico), Heidi K. Thornquist (Sandia National Laboratories), and Juliane Mueller (Lawrence Berkeley National Laboratory) discussed their past and present experiences with national laboratories. Mueller enjoys working with many domain scientists from different research areas, while Thornquist appreciates the diversity of her colleagues' backgrounds and the broad selection of challenging applications. "I feel inspired to come to work because you learn something new every day," she said. Schroder, who has spent most of his career in national laboratories, likes contributing to projects with far-reaching impact and collaborating with experts to apply their specific knowledge base to real-world problems.

Graduate students can focus on developing several skill sets in preparation for potential careers at a national lab. Because laboratory staff frequently utilize large machines, employers value good parallel computing skills and software professionalism. To avoid feeling overwhelmed when interacting with the many domain scientists at national laboratories, students should listen to other types of lectures during graduate school and attend seminar series outside of their departments. "Be able to admit that there are things you do not know, and be open-minded to learning from other people," Mueller said.

For researchers at national labs with a continuing passion for education, teaching opportunities are still available — though they often take a different form than in academia. "There are a lot of opportunities to work with students over the summer, get involved in their Ph.D.s, and serve on their committees," Schroder said.

Unique funding considerations exist for junior researchers who are establishing themselves at national laboratories. Many labs have early-career funds that are specifically meant for people to begin building their careers, though demand for these funds admittedly outstrips supply. When Mueller began working at Lawrence Berkeley, she discovered research opportunities by reaching out to colleagues who would benefit from her expertise in optimization. Since optimization is useful in a multitude of applications, she was able to connect with scientists during calls for funding opportunities and get involved with multiple projects. "Once you get the ball rolling, it is a lot easier," Mueller said. "People now reach out to me to do optimization on their projects."

Junior researchers might sometimes worry about becoming pigeonholed early in their careers. However, many areas of applied mathematics are valuable in numerous fields and are thus less niche. "Look for areas where your expertise is useful," Thornquist said. "This will guide you onto a path that keeps you from being pigeonholed." Asking for conversations and making connections can also lead to one's involvement in a variety of fields. "Reach out into a couple different areas," Schroder added. "Groups are very excited for people to sit in at their meetings."

Careers in Industry

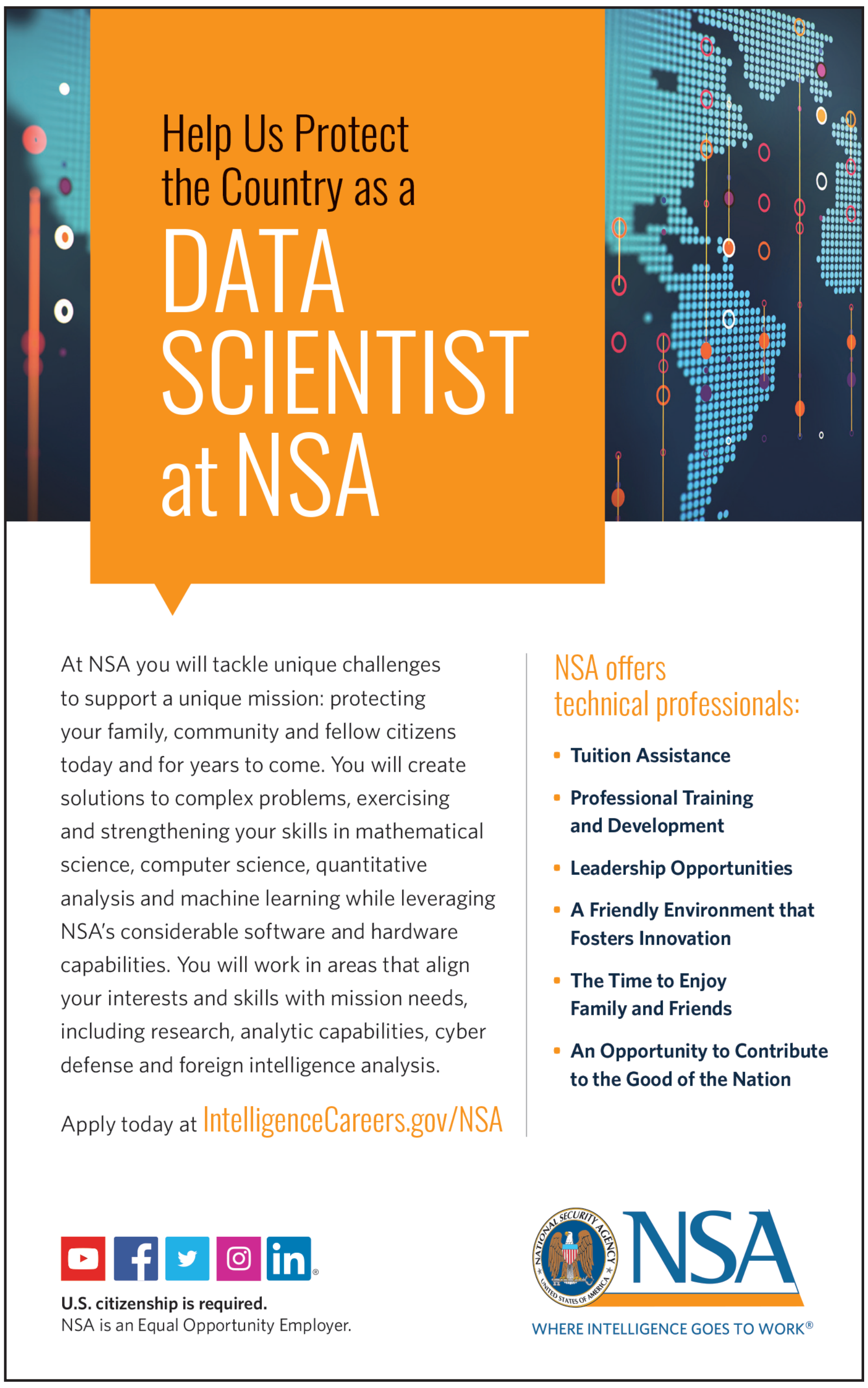
Julia Ling (Citrine Informatics), Maxim Naumov (Facebook), and Jeff Hammond (NVIDIA) explained that teamwork, the ability to receive feedback, and an aptitude for collaborative code development

See **Early Careers** *on page 12*

---

[1] https://www.siam.org/conferences/cm/conference/cse21

[2] https://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=72033



Professional Opportunities and Announcements

Send copy for classified advertisements and announcements to ***marketing@siam.org****.*
For rates, deadlines, and ad specifications, visit ***www.siam.org/advertising****.*

*Students (and others) in search of information about careers in the mathematical sciences can click on "Careers" at the SIAM website (****www.siam.org****) or proceed directly to* ***www.siam.org/careers****.*

Two Approaches to a Proof of Goldbach's Conjecture

Goldbach's Conjecture, which was announced in 1742, asserts that each positive even integer greater than 2 is the sum of two prime integers. Thus, e.g., $12 = 5 + 7$. The Conjecture is still unproved.

I believe that I have discovered two approaches to a proof — they are based on a strategy that has been effective with two other very difficult problems on which I have been working. It consists of finding a structure that contains all possibilities, and shows crucial relationships between them. As far as I know, the approach is original.

I will welcome comments. See the section referenced on the first page of the *second* part of "A Few Off-the-Beaten-Track Observations…" on **occampress.com**.

— *Peter Schorer,* peteschorer@gmail.com

The Future of Deep Learning Will be Sparse

By Torsten Hoefler, Dan Alistarh, Nikoli Dryden, and Tal Ben-Nun

Deep learning continues to deliver surprising new capabilities for tasks such as image and object classification, game play, translation, and even molecular structure prediction and other significant advancements in the computational sciences. These tasks are often carried out at human or superhuman performance levels, and one could argue that machines take over where human understanding of complex systems ends. The three key ingredients of deep neural networks (DNNs) are data, compute, and models; the latter includes the algorithms that design and train model structure and weights. In fact, the development of more capable deep learning systems is largely fueled by increasingly greater model sizes and amounts of data. OpenAI predicts an exponential growth in model size that appears to be supported by the data; between 2012 and 2018, training efforts grew by 300,000x.[1] Present-day researchers continue to employ scaling with trillion-parameter models. Yet although today's large models provide excellent performance, they seem relatively inefficient and are expensive to evaluate.

Over the last 10 years, both Moore's law and innovations in high-performance computing—such as graphics processing unit (GPU) accelerators—have driven the first wave of progress in deep learning. With the end of Moore's law now in sight, the "free" cost reductions in computation and storage to which we are accustomed will likely grind to a halt in the near future. However, inspiration from biology could lead to algorithmic solutions that pertain to the original neural networks from the 1960s. For instance, if we draw an analogy to biological brains—which are many orders of magnitude more energy efficient for similar tasks—we find that their connectivity is rather sparse. In fact, animal brains tend to become sparser as they increase in size. If we circle back to the aforementioned key ingredients, more data will remain available but more compute may break the bank. A growing body of research is investigating approaches that engineer sparsity into deep learning models to fuel the next decade of success in artificial intelligence. Existing works indicate that sparsity can achieve speedups between 10 and 100x in the near future; even more may be possible at higher sparsity levels.

But what exactly is "sparsity" in the context of deep learning? This is a surprisingly complex topic, and more than 300 research papers have described different techniques that address several aspects of the following fundamental questions: What should we sparsify? How can we sparsify? When should we sparsify? How can we integrate sparsity into training?

Sparsification comes in many shapes and forms, and we can broadly categorize them into two classes: model sparsity and ephemeral sparsity. In model sparsity, the sparse structure is related to the model itself; we can remove neurons, weights, or even whole substructures like filters or heads. In contrast, ephemeral sparsity is tied to the training or inference process and changes with each input example. This type of sparsity is well-known outside the realm of performance in operators such as Dropout and rectified linear units, where it prunes connections and activations respectively. Figure 1 depicts an overview and coarse classification of the various sparsity approaches in use today.

While sparsity reduces the number of arithmetic operations to be executed and weights to be stored, it incurs additional control and storage overheads that represent the sparse structure. In scientific computing, we would traditionally not even consider sparsity below 99 percent to be efficient. However, use of specialized architectures ensures that even 50 percent sparse workloads can yield performance benefits; NVIDIA's Ampere microarchitecture offers to multiply matrices with 50 percent sparsity nearly twice as quickly. Today's DNN models achieve sparsity ranges from 50 to 95 percent without significant accuracy loss, and ephemeral sparsity may yield additional savings. Yet we still do not know whether block-wise sparsity, which requires significantly lower representation and control overheads, can yield much of the benefits of fine-grained sparsity. Experiments show that the sparsity-accuracy tradeoff—which we define as "parameter efficiency" [2]—is lower, but the computational benefits may outweigh this loss. In practice, these benefits depend on both the problem and target computer architecture.

We must still examine many additional topics to gain a full picture of sparsity in deep learning. For example, how do we select elements for removal? Techniques range from simple "leave one out and check the quality" approaches, various saliency measures, and learned gating functions and regularizations to selection schemes based on linear or quadratic models of the loss function. If we sparsify during training, we may need to regrow other elements after removal to maintain a balance. We can do so randomly—based on the loss function's gradient—or with preferential attachment rules that are inspired by the brain and power-law graphs. Ephemeral sparsity approaches can lead to substantial memory and communication savings, ultimately yielding significant speedups with the right runtime support for sparse parallel reductions [3]. Our full overview paper on this subject provides an extensive outline of the details of sparse methods for deep learning [2].

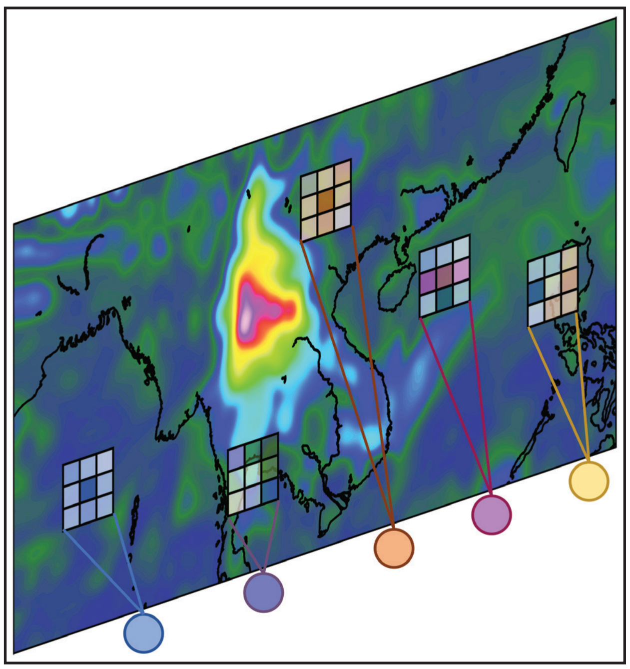


Figure 2. *Locally connected layers employ different weights for each region. Figure adapted from [1].*

DNNs are also beginning to gain popularity in many, if not most, scientific domains. Here, sparsity could be important even when exploring model options because it often leads to quality improvements in the low sparsity regime. For example, post-processing of weather and climate data can take advantage of locally connected layers to correct for biases in numerical model predictions for different parts of the Earth [1]. In a locally connected layer, each output neuron is only connected to a spatially constrained set, much like a stencil computation. In contrast, a fully connected layer—which connects each input neuron to every output neuron—exploits the domain's physically induced sparsity. The difference with convolutional layers that researchers use for image recognition is that the weights are not the same for all spatial points and can thus learn specific properties of each spatial region (see Figure 2). For example, the function that one applies on top of the ocean can differ from the function that one applies on top of a continent. We expect that such sparse techniques will become more relevant in deep learning for scientific computing and physical systems. As such, scientists should embrace them from the very beginning.

The future is bright for sparse deep learning; we will clearly begin to increase sparsity in the very near future, and most vendors are working on architectural support for their accelerators or software pipelines. Yet a multitude of challenges remain. How sparse can we go? Can we train with full sparsity? How can we understand sparsity's power? Furthermore, early research indicates that sparse neural networks may amplify existing biases and compromise fairness. But despite the many enduring complicated problems, sparsity in deep learning will surely rise in practical systems.

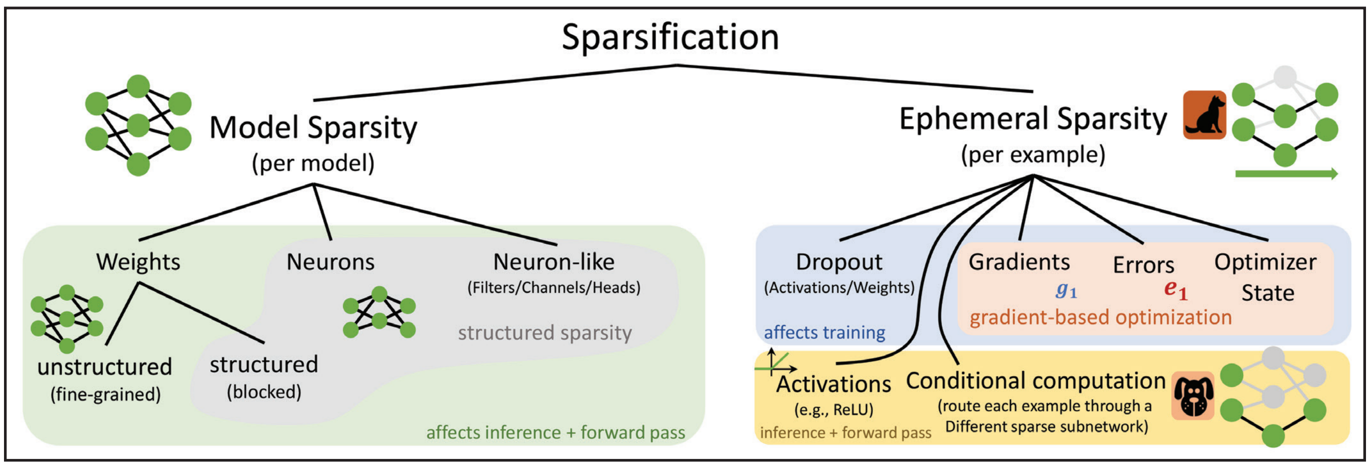


Figure 1. *An overview of sparsification approaches in deep learning. Figure courtesy of [2].*

References

[1] Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Phil. Trans. Roy. Soc. A.*, *379*(2194).

[2] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Preprint, *arXiv:2102.00554*.

[3] Renggli, C., Ashkboos, S., Aghagolzadeh, M., Alistarh, D., & Hoefler, T. (2019). SparCML: High-performance sparse communication for machine learning. In *SC '19: Proceedings of the international conference for high performance computing, networking, storage and analysis*. New York, NY: Association for Computing Machinery.

Torsten Hoefler[2] is a professor of computer science at ETH Zürich. His main research interests lie in large-scale supercomputing and networking, as well as deep learning systems. Dan Alistarh[3] is an assistant professor at the Institute of Science and Technology Austria and the machine learning research lead at Neural Magic, Inc. His research focuses on efficient algorithms that enable scalable deep learning training and inference. Nikoli Dryden[4] is an ETH Postdoctoral Fellow at ETH Zürich. His research focuses on high-performance deep learning and machine learning for science. Tal Ben-Nun is a postdoctoral fellow at ETH Zürich. His research interests include large-scale machine learning for scientific computing, learnable representations of code, and high-performance programming models.

---

[1] https://openai.com/blog/ai-and-compute

[2] http://htor.inf.ethz.ch

[3] https://people.csail.mit.edu/alistarh

[4] https://ndryden.com

Early Careers

Continued from page 11

are all necessary skills for industry careers. Applied mathematicians in the industry sector must also be able to converse with a wide range of people in various business roles. "Sometimes you have to convince people that they need to understand a topic," Hammond said. "You will have to be able to justify the money that you spend."

Some industry jobs are available to individuals without graduate degrees. Ling, who works at a start-up company, noted that graduate school is not essential for many roles due to the current start-up culture. But Naumov's group at Facebook strongly prefers employees to have a Ph.D. The panelists indicated that if the name of an organization or branch contains "labs" or "research," it likely has a culture that values Ph.D.s. Some companies even offer postdoctoral research positions that tend to pay better than similar postings in academia. "These positions can be very valuable and help guide your research towards practical, applied problems," Naumov said.

Though it might be challenging, it is possible to transition from industry to academia or a national laboratory. "The most important thing is just to be excellent," Ling said. "If you are excellent at your job, you will be recognized, and other people will want you to work for them." Maintaining strong connections with colleagues who can vouch for the quality of one's work is especially important when undergoing a large career shift, such as from industry to academia. Managers are taking a chance if they hire someone who is making this move, so they will value references very highly.

The panelists concluded by describing parts of their early careers that they particularly appreciate. Hammond has enjoyed the many interesting possibilities that his industry career enabled him to explore, while Naumov is glad that he ventured out of his comfort zone and learned new techniques from different areas and adjacent fields. Ling noted that she had not intended to conduct postdoctoral research, but seized the chance when it arose. She reminded attendees to be flexible. "It's okay to take an opportunity that is not a part of your plan," she said.

CSE21 attendees can access a recording of this panel and other conference presentations by logging in to the virtual platform. Recordings are available until June 5th. If you weren't able to attend the meeting, SIAM has reopened registration at a discounted rate for on-demand viewing of all talks. Learn more and register online.[3]

Jillian Kunze is the associate editor of SIAM News.

---

[3] https://www.siam.org/conferences/cm/registration/cse21-registration