Control Systems Reproducibility Challenge

The reproducibility of research was one of the main topics at the 2018 Panel of Editors meeting held in Los Angeles this past April. Richard Braatz, the previous editor-in-chief of *IEEE Control Systems Magazine*, discussed the concern of reproducibility of research in the broader field of computational research, leading to the question in [1] of "Should authors in the control field be expected or compelled to make their software public, as a way to reduce errors and to facilitate progress in the field?" Two replies were received [2], [3], both of which supported higher levels of reproducibility, and I also strongly support moving in that direction as well. The question remains of how best to achieve it.

(From left): Jonathan How, Edwin Chong, and Alessandro Astolfi enjoy a walk around Lake Como (courtesy of Li-Chen Fu).

Reference [1] discussed the initial efforts by Ian Mitchell and others within the context of the Association for Computing Machinery (ACM) International Conference on Hybrid Systems: Computation and Control (HSCC). Those efforts have continued, leading to a repeatability evaluation (HSCC-RE) [4] that is now available. The description of the HSCC-RE notes the dilemma that most authors don't post their code. Even for those who do, it is not easy to know if others can use that code to reproduce the results. Thus, computational results quickly become "nonreproducible—even by the research group that originally produced them." The goal of HSCC-RE is to improve the reproducibility of computational results in the papers selected. To participate, authors must create/submit a repeatability package. In response, they receive a repeatability evaluation package that rates the reproducibility of the work and can be linked/cited.

Our recent publication [5] underwent a similar evaluation wherein the various benchmark problems were retested, the unit tests for the major algorithms were run and evaluated, and the code itself was evaluated by eye and using automated tools such as *pylint* [6]. Thus, more than just reading over the article and accepting that the results were accurate as given, the reviewers dug deeply into the submission package and tested the foundation of the claims. As is often the case with that type of feedback, the reviewers found many issues that we overlooked and substantially improved the final product [7].

Creating and evaluating these repeatability packages can take a lot of extra time and effort. So how does this scale to conferences such as the American Control Conference (ACC) and IEEE Conference on Decision and Control (CDC)? Similar concerns exist in the machine learning community, but the organizers of the workshop [8] note that "part of ensuring reproducibility of the state-of-the-art is ensuring fair comparisons, proper experimental procedures, and proper evaluation methods and metrics."

Reference [9] presents an innovative solution approach by posing a "Reproducibility Challenge" for empirical results submitted to the 2018 International Conference on Learning Representations (ICLR). In this case, "reviewers" select an ICLR submission with the aim of replicating the experiments and determining if they are reproducible and support the conclusions of the paper. Thus, participants act as an inspector attempting to

Digital Object Identifier 10.1109/MCS.2018.2829978
Date of publication: 18 July 2018

verify the validity of the paper. However, since ICLR is an open-review process [10], that role can extend to helping the authors improve the quality of their work and paper (and even authorship). Feedback is provided as a reproducibility report that is more than a simple pass/fail—it will identify the parts of the contribution that can be reproduced and at what cost, in terms of resources (computation, time, people, development effort, and communication with the authors). Note that the new partnership between the IEEE and Code Ocean [11] provides an easy-to-use web platform in which users can share and run code in the cloud. Users are able to easily upload their code and associated data to the site, where other users are able to run and/or modify them.

Key to this process is that the target participants are the many instructors teaching graduate-level machine learning courses who are encouraged to use this reproducibility challenge as the final project of their course. To date, 12 courses are listed as participating, from schools such as McGill University, the University of Michigan, Princeton University, Tel Aviv University, and the University of California, Irvine. A nominal allowance of Google Cloud credits is provided to ensure that sufficient resources are available to the participants.

ACC and CDC are not open review (yet), so the feedback cannot be provided as conveniently. However, it seems like an excellent idea to enlist the large number of students taking graduate-level classes on control systems to evaluate the reproducibility of the ACC/CDC papers. Similar to HSCC, the outcome could be a linkable repeatability report that serves as a "badge of distinction" for the authors that their work can be reproduced by others (see [12] for an example of how the badging process is implemented in the ACM digital library).

This might just seem like a lot of extra work, and it is unclear if it is really scalable [13], but at least starting the process has the potential to substantially improve the quality/impact of the control work being published. I strongly recommend that something similar to a "Control Systems Reproducibility Challenge" be created and implemented. I look forward to your comments and reading about the future efforts to create such a challenge and improve the level of reproducibility in IEEE Control Systems Society published research.

Jonathan P. How

REFERENCES

[1] R. D. Braatz, "Reproducible research [from the editor]," *IEEE Control Syst.*, vol. 34, no. 4, pp. 6–7, Aug. 2014.

[2] D. Atherton, "[Feedback]," *IEEE Control Syst.*, vol. 35, no. 1, p. 19, Feb. 2015.

[3] T. P. Pavlic, "[Feedback]," *IEEE Control Syst.*, vol. 35, no. 3, pp. 19–20, June 2015.

[4] Repeatability evaluation. Accessed on: Apr. 8, 2018. [Online]. Available: https://www.hscc2018.deib.polimi.it/repeatability-evaluation

[5] A. Geramifard, C. Dann, R. H. Klein, W. Dabney, and J. P. How, "RLPy: A value-function-based reinforcement learning framework for education and research," *J. Mach. Learn. Res.*, vol. 16, pp. 1573–1578, Aug. 2015.

[6] Pylint. Accessed on: Apr. 8, 2018. [Online]. Available: https://www.pylint.org/

[7] RLPy reinforcement learning framework. Accessed on: Apr. 8, 2018. [Online]. Available: https://github.com/rlpy/rlpy

[8] Reproducibility in ML Workshop, ICML'18. Accessed on: Apr. 8, 2018. [Online]. Available: https://sites.google.com/view/icml-reproducibility-workshop/home

[9] ICLR 2018 reproducibility challenge. Accessed on: Apr. 8, 2018. [Online]. Available: http://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html

[10] J. P. How, "Open review process," *IEEE Control Syst.*, vol. 37, no. 3, pp. 3–5, June 2017. doi: 10.1109/MCS.2017.2675018.

[11] IEEE. Code ocean FAQs. Accessed on: Apr. 8, 2018. [Online]. Available: https://ieeexplore.ieee.org/Xplorehelp/#/faqs/code-ocean

[12] ACM. Artifact review and badging. Accessed on: Apr. 8, 2018. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-badging

[13] J. Bailleul, L. Hall, J. M. F. Moura, S. Hemami, G. Setti, M. Forster, G. Grenier, F. Zappulla, and J. Keaton. (2016, Nov.). Report on the first IEEE Workshop on the future of research curation and research reproducibility. Accessed on: Apr. 8, 2018. [Online]. Available: https://www.ieee.org/content/dam/ieee-org/ieee/web/org/ieee_reproducibility_workshop_report_final.pdf

Maxwell's Demon

Entropy has been defined at least twice in the history of science. First, it was defined in physics as thermodynamic entropy by Boltzmann (1872) and Gibbs (1878), and later it was defined in mathematics by Shannon (1948). Shannon's *information entropy* is a measure of information, whereas *thermodynamic entropy* is a measure of the number of states a physical system (like a jar of gas) can adopt.

These two different conceptualisations of entropy do not seem to be obviously related. But they are, and the relationship between them matters because thermodynamic entropy can be used to measure the energy cost of Shannon's information entropy. If this were not true then it would be possible to use a hypothetical being, known as *Maxwell's demon*, to run power stations on pure information.

—James V. Stone, *Information Theory—A Tutorial Introduction*,
Sebtel Press; 1st edition (February 1, 2015), page 171, 978-0956372857.