Date: 5/1/2023

R E S U M E

1. PERSONAL DETAILS

Full Name: Daniel Soudry
Identity No: 039933585
Date and place of birth: 5.9.1983, Haifa, Israel
Marital status: Married + 2 children
Phone numbers: 0507-989799
E-mail: daniel.soudry@technion.ac.il
Website: https://sites.google.com/site/danielsoudry/

2. ACADEMIC DEGREES

2008-2013 PhD, direct track
Electrical Engineering, Technion, Israel (Advisor: Prof. Ron Meir)
2004-2008 BSc, Summa cum Laude
Electrical Engineering and Physics, Technion, Israel

3. ACADEMIC APPOINTMENTS

2021– Present Associate Professor
Electrical Engineering, Technion, Israel
2017– 2021 Assistant Professor
Electrical Engineering, Technion, Israel
2014-2017 Post-doc
Statistics, Columbia University, USA (Host: Prof. Liam Paninski)

4. RESEARCH INTERESTS (*briefly*)

Despite the impressive recent progress using Artificial Neural Nets (ANNs), they are still far behind the capabilities of biological neural nets in most areas: even the simplest fly is far more resourceful and efficient than our most advanced autonomous systems. With the long-term aim of closing this gap, my research focuses on theoretically understanding how ANNs learn and operate, and how they can be improved, especially "opening up" computational bottlenecks.

For example, one bottleneck is the fragility of ANNs: even small changes to their training procedure can cause large degradation in their performance on test data, and it is not clear why. We developed several theoretical results to elucidate this phenomenon, relating it to the "implicit bias" hidden in the training algorithm, and how it helps find "good" (e.g., simple) solutions. This helped explain various puzzling practical observations.

Another bottleneck is the long time required to train large ANNs. A popular and simple method for accelerating ANNs training is to parallelize each optimization iteration by splitting the data on multiple workers. Previous works suggested this causes an inherent degradation in generalization. However, we showed the contrary: by correctly adjusting the algorithm and its number of steps this degradation can be avoided. This later led to significant speed-ups.

Another bottleneck is the energy consumption of ANNs. This can be improved significantly by reducing the numerical precision of these models and their training methods. Indeed, nearly all recent deep learning related hardware accelerators rely heavily on lower precision math. we showed how to reach very low precision (e.g., as low as 1bit for the model, or 4bit for the training method).

To further improve the resource efficient of ANN, in the future we aim co-optimize the training algorithm together with the hardware required to train it, perhaps even directly learning the best circuit implementing the ANN (e.g., growing the hardware implementation of the ANN directly from the silicon). This interdisciplinary direction would require collaboration with other researchers from the area of semiconductors or circuit design.

5. TEACHING EXPERIENCE

Undergraduate courses: "Machine Learning", "Biological Signals and Systems", "Deep Learning"
Graduate courses: "Theoretical Topics in Deep Learning"

7. DEPARTMENTAL ACTIVITIES

2021-Present Member of the Faculty's undergraduate studies committee
2018–2020 Member of the Faculty's committee for students with borderline academic status
2017–Present Advisor for undergraduate students in Machine learning
2017–Present Organizing ML, Control and Systems area in undergraduate and graduate studies

8. PUBLIC PROFESSIONAL ACTIVITIES

2019–Present Area chair at ICML (2019, 2020), NeurIPS (2020,2021), ICLR (2021,2022)
2017–Present Grant Reviewer at ISF, BSF, and Swiss Science Foundation.
2010–Present Reviewer at various top-tier conferences (NeurIPS, ICML, ICLR, COLT) and journals (e.g., JMLR, IEEE TNNLS/ TPMI/TNANO, Physical review Letters/E/X).

2019-present Contact with Intel in the Technion ML center and reviewing research proposals.

9. FELLOWSHIPS, AWARDS AND HONORS

Awards as faculty member

2022-2025 Schmidt Career Advancement Chair in Artificial Intelligence (8,333$ per year to salary, and 25,000$ per year as research gift money).
2020 Rising Star Faculty Award (30000$ research gift money), one of ten winners selected from all over the world (and the only one outside of the USA).
2020 Alexander Goldberg Research Prize (3000$), for research done on "Resource-Efficient Deep Learning", and contributions to the Israeli industry.
2017–2019 Taub Fellowship for leaders in science and technology, Technion, Israel

Post-doc Fellowships

2014-2016 Gruss-Lipper Post-Doctoral Fellowship (190890$), Gruss Lipper Charitable foundation, USA
2013 MIT-Technion Post-Doctoral Fellowship (declined), MIT and Technion, USA

Student Awards

2012 Jury Award for excellent graduate students, Technion, Israel
2012 Jacobs excellence scholarship for graduate students, Technion, Israel
2010 Zeff excellence scholarship for graduate students, Technion, Israel
2009 Sherman excellence scholarship for graduate students, Technion, Israel
2008 Tzifers award for excellent new graduate students, Technion, Israel
2008 Fintzi award for excellence in BSc studies, Technion, Israel
2008 Summa cum Laude BSc, with GPA 96/100.

10. GRADUATE STUDENTS

Completed PhD theses

- Elad Hoffer, completed in 2019, "Deep Learning: Rethinking Common Practices", co-supervisor: Nir Ailon. Currently works as a research lead at Habana.ai.
- Edward Moroshko, completed in 2021, "On Implicit Bias in Deep Models and Constrained Feedback in Online Learning", previous supervisor: Koby Crammer. Currently works with me as a post-doc associate.

- Itay Hubara, completed in 2022, "Towards Fast and Efficient Deep Learning". Currently works as a research lead at Habana.ai.

Completed MSc theses

- Chen Zeno, completed in 2019, "Task Agnostic Continual Learning Using Online Variational Bayes". Currently a PhD student in my group.
- Itay Golan, completed in 2020, "Effects of human-controlled hyper-parameters in deep neural networks", works at Final.
- Liad Ben-Ori, completed in 2021, "Improving Efficiency of DNN Training Using Stochastic Pruning".

PhD theses in progress

- Mor Shpigel-Nacson, started on 2017 (direct route), expected graduation: 2022
- Niv Giladi, started on 2017 (direct route), expected graduation: 2022. Also works at Habana.ai.
- Chen Zeno, started on 2019, expected graduation: 2023
- Itay Evron, started on 2019, expected graduation: 2023
- Yaniv Blumenfeld, started on 2019 (direct route), expected graduation: 2024
- Matan Haroush, started on 2020 (direct route), expected graduation: 2025

- Tzofnat Grinberg, I became her co-supervisor on 2019 (Primary supervisor: Shahar Kvatinsky).
- Brian Chmiel, I became his co-supervisor on 2020 (Primary supervisor: Alex Bronstein).

MSc theses in progress

- Hagay Michaeli, started on 2021, expected graduation: 2023
- Shahar Gottlieb, started on 2022, expected graduation: 2024

11. SPONSORED LONG-TERM VISITORS AND POST-DOCTORAL ASSOCIATES

- Edward Moroshko, post-doc associate, started 2020.
- Alon Bruzkus, post-doc associate, started 2022.

12. RESEARCH GRANTS

Competitive

Project Title	*Funding source*	*Amount (USD)*	*Period*	*Role of the PI*
A-B-C-Deep: Algorithmic Bias Control in Deep Learning	European Research Council	1,647,360	2022 -2027	Sole PI
Towards Massively Parallel and Resource Efficient Deep learning: Theory and Practice	Israeli Science Foundation	241,908	10/1/2018-30/9/2022	Sole PI

Industrial and other sources

Project Title	*Funding source*	*Amount (USD)*	*Period*	*Role of the PI*
Flexpoint Training:	Intel	250,000	1/1/2018-	Sole PI

Theory and Practice (4 separate grants)			31/12/2022	
Low numerical precision in resource constrained neural networks	Israel Innovation Authority	225,409	1/8/2019-31/8/2022	PI in a MAGNET Consortium
Evaluators of Deep Neural Networks	AIGrant.org	2500 +20,000 GPU credits	2017	Sole PI

13. PUBLICATIONS

Note: In the papers below, I underscored the author names of students (or who were my students at the time the paper was written/published), for publications after I became a faculty member.

Theses

- Thesis title: "Neurons: from Biophysics to Functionality", Supervisor: Ron Meir, degree received: December 18th, 2013.

Refereed papers in professional journals

1. **D. Soudry** and R. Meir, "History-Dependent Dynamics in a Generic Model of Ion Channels–An Analytic Study", Front. Comput. Neurosci., vol. 4, Jan. 2010. [Impact Factor 2.653].
2. **D. Soudry** and R. Meir, "Conductance-based neuron models and the slow dynamics of excitability", Front. Comput. Neurosci., vol. 6, no. 4, 2012. [Impact Factor 2.653].
3. P. Orio and **D. Soudry**, "Simple, fast and accurate implementation of the diffusion approximation algorithm for stochastic ion channels with multiple states", PLoS ONE, vol. 7, no. 5 p. e36670, 2012. [Impact Factor 3.234].
4. **D. Soudry** and R. Meir, "The neuronal response at extended timescales: long term correlations without long memory". Front. Comput. Neurosci., vol. 8, no. 35, 2014. [Impact Factor 2.653].
5. **D. Soudry** and R. Meir, "The neuronal response at extended timescales: a linearized spiking input-output relation". Front. Comput. Neurosci., vol. 8, no. 29, 2014. [Impact Factor 2.653].
6. D. Pezo, **D. Soudry**, P. Orio, "Diffusion approximation-based simulation of stochastic ion channels: which method to use?". Front. Comput. Neurosci., vol. 8, no. 139, 2014 (Part of the research topic "Neuronal stochastic variability: influences on spiking dynamics and network activity"). [Impact Factor 2.653]
7. **D. Soudry**, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Memristor-based multilayer neural networks with online gradient descent training". IEEE TNNLS, vol. 26, no. 10, 2015. [Impact Factor 4.37].
8. **D. Soudry**, S. Keshri, P. Stinson, M.H. Oh, G. Iyengar, L. Paninski, "Efficient 'Shotgun' Inference of Neural Connectivity from Highly Sub-sampled Activity Data", PLoS Comput Biol, vol. 11 no. 10, 2015. [Impact Factor 4.62].
9. E. A. Pnevmatikakis, **D. Soudry**, Y. Gao, T. A. Machado, J. Merel, D. Pfau,T. Reardon,Y. Mu, C. Lacefield, W. Yang, M. Ahrens, R. Bruno, T. M. Jessell, D. S. Peterka, R. Yuste, L. Paninski, "Simultaneous Denoising, Deconvolution, and

Demixing of Calcium Imaging Data", Neuron, vol. 89, no. 2, 2016. [Impact Factor 13.974].

10. Y. Dordek*, **D. Soudry***, R. Meir, D. Derdikman (*contributed equally), "Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis", eLife, vol. 5, e10094, 2016. [Impact Factor 8.303, F1000 Recommended].
11. J. Friedrich, W. Yang, **D. Soudry**, Y. Mu, M. B. Ahrens, R. Yuste, D. S. Peterka, L. Paninski, "Multi-scale approaches for high-speed imaging and analysis of large neural populations", PLoS Comput Biol, vol., 13 no. 8, e1005685, 2017. [Impact Factor 4.62]
12. S. Ahmadizadeh, P. Jane Karoly, D. Nesic, D. Br. Grayden, M. J.Cook, **D. Soudry**, D. R. Freestone, "Bifurcation Analysis of Two Coupled Jansen-Rit Neural Mass Models", PLOS One, vol. 13 no. 3, e0192842, 2018. [Impact Factor 3.54]
13. I. Hubara*, M. Courbariaux*, **D. Soudry**, R. El-Yaniv, and Y. Bengio (*equal contribution), "Quantized neural networks: Training Neural Networks with Low Precision Weights and Activations", JMLR 2018. [Impact Factor 2.45].
14. P. J. Karoly, L. Kuhlmann, **D. Soudry**, D. B. Grayden, M. J. Cook, D. R. Freestone, "Seizure pathways: A model-based investigation", PLoS Comput Biol, 2018.
15. **D. Soudry**, E. Hoffer, M. Shpigel Nacson, S. Gunasekar, N. Srebro, "The Implicit Bias of Gradient Descent on Separable Data", JMLR, 2018.
16. Z. Zhu, **D. Soudry**, Y. C. Eldar, M. B. Wakin, "The Global Optimization Geometry of Shallow Linear Neural Networks", Journal of Mathematical Imaging and Vision, 2019.
17. C. Zeno*, I. Golan*, E. Hoffer, **D. Soudry**, "Task Agnostic Continual Learning Using Online Variational Bayes with Fixed-Point Updates" (*equal contribution), Neural Computation, 2021.
18. T. Greenberg-Toledo, B. Perach, I. Hubara, **D. Soudry**, S. Kvatinsky, "Training of Quantized Deep Neural Networks using a Magnetic Tunnel Junction-Based Synapse", Semiconductor Science and Technology, 2021.

Refereed papers in conference proceedings

1. D. B. Chklovskii and **D. Soudry**, "Neuronal spike generation mechanism as an oversampling, noise-shaping A-to-D converter", NIPS 2012 [25.2% acceptance rate].
2. **D. Soudry**, I. Hubara and R. Meir, "Expectation Backpropagation: Parameter-Free Training of Multilayer Neural Networks with Continuous or Discrete Weights", NIPS 2014. [24.7% acceptance rate].
3. S. Greshnikov, E. Rosenthal, **D. Soudry**, and S. Kvatinsky, "A Fully Analog Memristor-Based Multilayer Neural Network with Online Backpropagation Training", IEEE ISCAS 2016 [acceptance rate unknown, but was 27.18% in 2015], pp. 1394-1397.
4. I. Hubara*, M. Courbariaux*, **D. Soudry**, R. El-Yaniv, and Y. Bengio (*equal contribution), "Binarized neural networks", NIPS 2016. [22.7% acceptance rate].
5. E. Hoffer***,*** I. Hubara*, **D.Soudry**, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks" (*equal contribution), NIPS 2017, **Oral presentation [1.2% acceptance rate]**.
6. E. Hoffer, I. Hubara, **D. Soudry**, "Fix your classifier: the marginal value of training the last weight layer", ICLR 2018.

7. **D. Soudry**, E. Hoffer, M. Shpigel Nacson, S. Gunasekar, N. Srebro, “The Implicit Bias of Gradient Descent on Separable Data”, ICLR 2018.
8. S. Gunasekar, J. Lee, **D. Soudry**, N. Srebro, “Characterizing Implicit Bias in Terms of Optimization Geometry”, ICML 2018.
9. R. Banner, I. Hubara, E. Hoffer, **D. Soudry**, “Scalable Methods for 8-bit Training of Neural Networks”, NIPS 2018.
10. S. Gunasekar, J. D. Lee, **D. Soudry**, N. Srebro, “Implicit Bias of Gradient Descent on Linear Convolutional Networks”, NIPS 2018.
11. E. Hoffer, R. Banner, I. Golan, **D. Soudry**, “Norm matters: efficient and accurate normalization schemes in deep networks”, NIPS 2018, **Spotlight Presentation [3.5% acceptance rate]**.
12. M. Shpigel-Nacson, N. Srebro, **D. Soudry**, “Stochastic Gradient Descent on Separable Data: Exact Convergence with a Fixed Learning Rate”, AISTATS 2019.
13. M. Shpigel-Nacson, J. Lee, S. Gunasekar, N. Srebro, **D. Soudry**, “Convergence of Gradient Descent on Separable data”, AISTATS 2019, **Oral Presentation [2.5% acceptance rate]**.
14. P. Savarese, I. Evron, **D. Soudry**, N. Srebro, "How do infinite width bounded norm networks look in function space?", COLT 2019.
15. M. Shpigel Nacson, S. Gunasekar, J. Lee, N. Srebro, **D. Soudry**, "Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models", ICML 2019.
16. R. Banner, Y. Nahshan, **D. Soudry**, "Post-training 4-bit quantization of convolution networks for rapid-deployment", NeurIPS 2019.
17. Y. Blumenfeld, D. Gilboa, **D. Soudry**, “A Mean Field Theory of Quantized Deep Networks: The Quantization-Depth Trade-Off”, NeurIPS 2019.
18. G. Ongie, R. Willett, **D. Soudry**, N. Srebro, "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case", ICLR 2020.
19. N. Giladi, M. Shpigel Nacson, E. Hoffer, **D. Soudry**, "At Stability's Edge: How to Adjust Hyperparameters to Preserve Minima Selection in Asynchronous Training of Neural Networks?", ICLR 2020, **Spotlight Presentation [4.1% acceptance rate]**.
20. E. Hoffer, T. Ben-Nun, N. Giladi, I. Hubara, T. Hoefler, **D. Soudry**, “Augment Your Batch: Improving Generalization Through Instance Repetition”, CVPR 2020.
21. M. Haroush, I. Hubara, E. Hoffer, **D. Soudry**, "The Knowledge Within: Methods for Data-Free Model Compression", CVPR 2020.
22. B. Woodworth, S. Gunasekar, P. Savarese, E. Moroshko, I. Golan, J. Lee, **D. Soudry**, N. Srebro, "Kernel and Deep Regimes in Overparametrized Models", COLT 2020.
23. Y. Blumenfeld, D. Gilboa, **D. Soudry**, “Beyond Signal Propagation: Is Feature Diversity Necessary in Deep Neural Network Initialization?”, ICML 2020.
24. E. Moroshko, S. Gunasekar, B. Woodworth, J. D. Lee, N. Srebro, **D. Soudry**, "Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy", NeurIPS 2020, **Spotlight Presentation [3% acceptance rate].**
25. B. Chmiel*, L. Ben-Uri*, M. Shkolnik, E. Hoffer, R. Banner, **D. Soudry**, "Neural gradients are near-lognormal: improved quantized and sparse training". (*Indicates equal contribution), ICLR 2021.
26. S. Azulay, E. Moroshko, M. Shpigel Nacson, B. Woodworth, N. Srebro, A. Globerson, **D. Soudry**, "On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent", ICML 2021, Long talk **[3% acceptance rate]**.

27. I. Hubara*, Y. Nahshan*, Y. Hanani*, R. Banner, **D. Soudry**, "Accurate Post Training Quantization with Small Calibration Sets", (*Indicates equal contribution) ICML 2021.
28. I. Hubara, B. Chmiel, M. Island, R. Banner, S. Naor, **D. Soudry**, "Accelerated Sparse Neural Training: A Provable and Efficient Method to Find N:M Transposable Masks", NeurIPS 2021.
29. N. Giladi, Z. Ben-Haim, S. Nevo, Y. Matias, **D. Soudry**, "Physics-Aware Downsampling with Deep Learning for Scalable Flood Modeling", NeurIPS 2021.
30. R. Mulayoff, T. Michaeli, **D. Soudry**, "The Implicit Bias of Minima Stability: A View from Function Space", NeurIPS 2021.
31. A. Tamar, **D. Soudry**, E. Zisselman, "Regularization Guarantees Generalization in Bayesian Reinforcement Learning through Algorithmic Stability", AAAI 2022 [15% acceptance rate].
32. M. Haroush, T. Frostig, R. Heller, **D. Soudry**, "A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks", ICLR 2022.
33. I. Evron, E. Moroshko, R. Ward, N. Srebro, **D. Soudry**, "How catastrophic can catastrophic forgetting be in linear regression?", COLT 2022.
34. M. Shpigel-Nacson, K. Ravichandran, N. Srebro, **D. Soudry**, "Implicit Bias of the Step Size in Linear Diagonal Neural Networks", ICML 2022.
35. M. Shpigel Nacson, R. Mulayoff, G.. Ongie, T. Michaeli, **D. Soudry**, "The Implicit Bias of Minima Stability in Multivariate Shallow ReLU Networks", ICLR 2023.
36. B. Chmiel, I. Hubara, R. Banner, **D. Soudry**, "Optimal Fine-Grained N:M sparsity for Activations and Neural Gradients", ICLR 2023 ("notable top 25%" of accepted papers).
37. B. Chmiel, R. Banner, E. Hoffer, H. Ben Yaacov, **D. Soudry**, "Logarithmic Unbiased Quantization: Practical 4-bit Training in Deep Learning", ICLR 2023.
38. I. Evron*, O. Onn*, T. Weiss, H. Azeroual, **D. Soudry**, "The Role of Codeword-to-Class Assignments in Error Correcting Codes: An Empirical Study", AISTAT 2023 (*Indicates equal contribution).

Patents granted.

1. I. Hubara, **D. Soudry**, and R. El-Yaniv, "Binarized Neural Networks", US Patent 10,831,444, Granted: 2020.
2. **D. Soudry**, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Analog Multiplier Using Memristor a Memristive Device and Methods for Implementing Hebbian Learning Rules Using Memristor Arrays", US Patent US9754203 B2, Granted: 2017.

16. CONFERENCES

Plenary, keynote or invited talks

- Invited talk at 3rd SLowDNN international workshop on "Seeking Low Dimensionality in Deep Neural Networks", at MBZUAI, Abu Dhabi, 3rd-6th of January 2023.
- Invited talk at the International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics), at King's College London, 17th-19th December 2022.
- Invited talk at the international "Conference on the Mathematics of Deep Learning" (DeepMath 2022), at University of California, San Diego, 17th-18th of November 2022.
- Keynote talk in "Machine Learning on HPC Systems" (MLHPCS) workshop, part of the International Supercomputing Conference (ISC), held online (2.7.2021)

- Invited talk at the “1st CVPR workshop on Binary Networks”, CVPR 2021, virtual event (25.6.2021).
- “Theory of Deep Learning: Where next?" Workshop, Institute for Advanced Study (IAS), Princeton, on 15-18.10.2019. (Invited to talk, but invitation declined due to childbirth)
- Invited talk at "Deep Learning and the Brain" International Symposium, ELSC, Hebrew University, Israel, on 20-22.1.2019 (invited talk). Most speakers where prominent speakers from abroad (and some from Israel).

Other invitations

- Participated in the Simons Institute Program on “Foundations of Deep Learning” (on 7/2019), Berkeley, which is a “by-invitation-only” event.
- “Theoretical aspects of machine learning” workshop, Tel-Aviv university, Israel, on 31.12.2019 (invited talk). This was a one-time event, most speakers were from Israel, and some prominent speakers from abroad.
- "AI week" conference, Tel-Aviv University, Israel, on 18-19.11.2019. This is an annual event with key speakers from the academia and the industry (mostly from Israel, some from abroad).
- Bosch AI CON, Bosch corporate research center in Renningen, Germany, on 19.11.2018. This is an annual event with key speakers from the academia and the industry.
- I have additionally been invited to speak in two colloquia, and more than 22 departmental seminars and industry events.

Contributed Talks and Posters

See conference proceedings (Oral or spotlight = talk, others=posters).

Participation in organizing conferences

“Deep Learning: Theory and Practice”, Technion, Israel, 6.6.2018, Co-Chair.