Levent Sagun

FAIR, Meta - Paris, France — leventsagun@gmail.com — Google Scholar — Home Page

Research Interests

I study failure modes in large models, focusing on contextualized measurement for AI governance including construct validity, brittleness, spurious correlations, and fairness under distribution shift. My work aims to build methods that reveal where models fail in context rather than their performance on benchmarks.

Professional Experience

2025 - today Senior Research Scientist, FAIR, Meta
Leading contextualized measurement research on AI governance at FAIR Alignment

2022 - 2024 Research Science Manager, FAIR, Meta
Led the Society & Responsible AI research area in EMEA and managed a research team
Drove value-driven evaluation and governance-aligned measurement directions

2019 - 2022 Research Scientist, FAIR, Meta
Led work on value-driven robustness and fairness under distribution shift
Built the Society & Responsible AI research pillar

2018 - 2019 Postdoctoral Fellow, École Polytechnique Fédérale de Lausanne, Switzerland
Simons Collaboration on Cracking the Glass Problem - Part II
Focused on “double-descent” and over-parametrization in deep learning with Matthieu Wyart

2017 - 2018 Postdoctoral Fellow, École Normale Supérieure Paris & CEA Saclay, France
Simons Collaboration on Cracking the Glass Problem - Part I
Researched algorithmic dynamics and optimization in deep learning with Giulio Biroli

2016 - 2017 Research Intern and Contractor, Facebook AI Research, New York
Studied optimization in deep learning and high dimensional loss landscapes with Léon Bottou

2015 - 2016 Teaching Assistant, Center for Data Science, New York University

2011 - 2017 Research & Teaching Assistant, Department of Mathematics, Courant Institute, NYU

Mentoring

2019 - today Mentored 10+ researchers at FAIR (6 PhD interns, 4 PhD students, 1 postdoc, 1 resident)
Directly supported and supervised 3 early-career researchers in Society and Responsible AI

Education

2011 - 2017 Ph.D. in Mathematics, Courant Institute of Mathematical Sciences, New York University
Thesis title: Explorations on High Dimensional Landscapes: Spin Glasses and Deep Learning
Academic advisors: Gérard Ben Arous, Yann LeCun, & Léon Bottou

2006 - 2011 B.S. in Mathematics and Physics, Boğaziçi University, Istanbul, Turkey
Exchange program: Columbia University, New York, Fall 2009
Graduated first rank in the Faculty of Arts and Sciences, Class of 2011

Papers

Fairness, Validity and Contextual Measurement (2022 - Present)

1. Issues in Measuring the Fairness of Social Representation in Synthetic (Speech) Data. Arjun Subramonian, Brooklyn Sheppard, **Levent Sagun**. Synthetic Data Workshop at Aarhus Decennial Conference 2025
2. Chantal Shaib, Vinith Suriyakumar, **Levent Sagun**, Byron Wallace, Marzyeh Ghassemi. Learning the Wrong Lessons: Syntactic-Domain Spurious Correlations in Language Models. Spotlight at NeurIPS 2025
3. Brooklyn Sheppard, Elia Ovalle, Adina Williams, **Levent Sagun**. On the lack of queer voices in diverse speech datasets. Social Science and Language Models workshop at Weizenbaum Institute 2025 & Speech AI for All Workshop at CHI 2025
4. Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Kai-Wei Chang, Adina Williams, **Levent Sagun**. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. Queer in AI at NeurIPS 2024 & FAccT 2025
5. Samuel Bell, Mariano Coria Meglioli, Megan Richards, Eduardo Sanchez, Christophe Ropers, Skyler Wang, Adina Williams, **Levent Sagun**, Marta Costa-jussa. On the role of speech data in reducing toxicity detection bias. Paper and dataset release at NAACL 2025
6. Krunoslav Lehman Pavasovic, David Lopez-Paz, Giulio Biroli, **Levent Sagun**. A Differentiable Rank-Based Objective For Better Feature Learning. ICLR 2025
7. Arjun Subramonian, Samuel J Bell, **Levent Sagun**, Elvis Dohmatob. An Effective Theory of Bias Amplification. ICLR 2025
8. **Levent Sagun**, Kartik Ahuja, Elvis Dohmatob, Julia Kempe. On generated vs collected data. The Workshop on Global AI Cultures at ICLR 2024
9. Arjun Subramonian, **Levent Sagun**, Yizhou Sun. Networked inequality: Preferential attachment bias in graph neural network link prediction. ICML 2024
10. Megan Ung, Alicia Sun, Samuel Bell, Bhaktipriya Radharapu, **Levent Sagun**, Adina Williams. Chained Tuning Leads to Biased Forgetting. Next Generation of AI Safety & TiFA Workshops at ICML 2024
11. Samuel Bell, **Levent Sagun**. Simplicity bias leads to amplified performance disparities. FAccT 2023
12. Vitor Albiero, Raghav Mehta, Ivan Evtimov, Samuel Bell, **Levent Sagun**, Aram Markosyan. Confusing Large Models by Confusing Small Models. OOD Workshop at ICCV 2023
13. Maximilian Nickel, Matthew Le and **Levent Sagun**. Diversity and Inclusion in Data Collection via Social Sampling. EAAMO 2022
14. Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, **Levent Sagun**, Nicolas Usunier. Fairness Indicators for Systematic Assessments of Visual Feature Extractors. FAccT 2022

Deep Learning Architectures, Optimization Dynamics and Over-parametrization (2015 - 2022)

1. Arjun Subramonian, **Levent Sagun**, Kai-Wei Chang, Yizhou Sun. Group Excess Risk Bound of Overparameterized Linear Regression with Constant-Stepsize SGD, Trustworthy and Socially Responsible Machine Learning Workshop at NeurIPS 2022
2. Stéphane d'Ascoli, Marylou Gabrié, **Levent Sagun**, Giulio Biroli. On the interplay between data structure and loss function in classification problems. OPPO ICML 2021 & NeurIPS 2021
3. Stéphane dAscoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, **Levent Sagun**. Convit: Improving vision transformers with soft convolutional inductive biases. ICML 2021

4. Stéphane d'Ascoli, **Levent Sagun**, Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? NeurIPS 2020

5. Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, **Levent Sagun**, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. Journal of Statistical Mechanics: Theory and Experiment 2020

6. Stéphane d'Ascoli, **Levent Sagun**, Joan Bruna, Giulio Biroli. Finding the Needle in the Haystack with Convolutions: on the benefits of architectural bias. NeurIPS 2019

7. Umut Simsekli, **Levent Sagun**, Mert Gurbuzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. ICML 2019

8. Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, **Levent Sagun**, Marco Baity-Jesi, Giulio Biroli, Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. **Phys. Rev. E 100**, 012115, 2019

9. Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, **Levent Sagun**, Giulio Biroli, Matthieu Wyart. A jamming transition from under- to over-parametrization affects loss landscape and generalization. Integration of Deep Learning Theories Workshop at NeurIPS 2018

10. Marco Baity-Jesi, **Levent Sagun**, Mario Geiger, Stefano Spigler, Gérard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, Giulio Biroli. Comparing Dynamics: Deep Neural Networks versus Glassy Systems. ICML 2018

11. David Lopez-Paz, **Levent Sagun**. Easing non-convex optimization with NNs. Workshop at ICLR 2018

12. **Levent Sagun**, Utku Evci, V. Uğur Güney, Yann Dauphin, Léon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. Workshop at ICLR 2018

13. **Levent Sagun**, Thomas Trogdon, Yann LeCun. Universal halting times in optimization and machine learning. Optimization Workshop Spotlight ICML 2016 & Journal pub. at Quart. Appl. Math. 2018

14. Andrew Ballard, Ritankar Das, Stefano Martiniani, Dhagash Mehta, **Levent Sagun**, Jacob Stevenson, David Wales. Perspective: Energy Landscapes for Machine Learning. Phys. Chem. Chem. Phys. 2017

15. P Chaudhari, A Choromanska, S Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, **Levent Sagun**, Riccardo Zecchina. Entropy-SGD: Biasing GD Into Wide Valleys. ICLR 2017

16. **Levent Sagun**, Uğur Güney, Gérard Ben Arous, Yann LeCun. Explorations on high dimensional landscapes. Workshop at ICLR 2015

Applied and Interdisciplinary Works

1. Matthew Le, Mark Ibrahim, **Levent Sagun**, Timothee Lacroix, Maximilian Nickel. Neural relational autoregression for high-resolution COVID-19 forecasting. epiDAMIK 4.0: The 4th International workshop on Epidemiology meets Data Mining and Knowledge discovery, 2021
Developed regional-relational forecasting models integrated into public health prediction pipelines.

2. **Levent Sagun**, Caglar Gulcehre, Adriana Romero, Negar Rostamzadeh, Stefano Sarao Mannelli. Post-Workshop Report on Science meets Engineering in Deep Learning, Workshop at NeurIPS 2019
Organizers as co-authors highlight emerging challenges and directions in deep learning research.

3. Matthew Dunn, **Levent Sagun**, Mike Higgins, Uğur Güney, Volkan Cirik, Kyunghyun Cho. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. arXiv:1704.05179, 2017
Developed a large-scale benchmark dataset for open-domain question answering from real web snippets.

4. Matthew Dunn, **Levent Sagun**, Hale Şirin, Daniel Chen. Early Predictability of Asylum Court Decisions. ML and the Law, **ICAIL**, 2017
Applied machine learning to reveal structural disparities and inconsistencies in asylum adjudication.

Preprints

1. Patrick Haller, Mark Ibrahim, Polina Kirichenko, **Levent Sagun**, Samuel Bell. LLM Knowledge is Brittle: Truthfulness Representations Rely on Superficial Resemblance. arXiv:2510.11905, 2025
2. Samuel Bell, Diane Bouchacourt, **Levent Sagun**. Reassessing the validity of spurious correlations benchmarks. arXiv:2409.04188, 2024
3. Arjun Subramonian, Adina Williams, Maximilian Nickel, Yizhou Sun, **Levent Sagun**. Weisfeiler and Leman Go Measurement Modeling: Probing the Validity of the WL Test. arXiv:2307.05775, 2023
4. David Lopez-Paz, Ishmael Belghazi, Diane Bouchacourt, Elvis Dohmatob, Badr Youbi Idrissi, **Levent Sagun**, Andrew Saxe. An Ensemble View on Mixup. OpenReview, 2023
5. Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, **Levent Sagun**, Armand Joulin, Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv:2202.08360, 2022
6. David Lopez-Paz, Diane Bouchacourt, **Levent Sagun**, Nicolas Usunier. Measuring and signing fairness as performance under multiple stakeholder distributions. arXiv:2207.09960, 2022
7. Berfin Simsek, Melissa Hall, **Levent Sagun**. Understanding out-of-distribution accuracies through quantifying difficulty of test samples. arXiv:2203.15100, 2022
8. Stéphane d'Ascoli, **Levent Sagun**, Giulio Biroli, Ari Morcos. Transformed CNNs: recasting pre-trained convolutional layers with self-attention. arXiv:2106.05795, 2021
9. Umut Simsekli, Mert Gurbuzbalaban, Thanh Huy Nguyen, Gael Richard, **Levent Sagun**. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. arXiv:1912.00018, 2019
10. **Levent Sagun**, Léon Bottou, Yann LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond. arXiv:1611.07476, 2017

Service

1. **Co-Organizer:** Society and Responsible AI seminar series at FAIR, Meta (2020 - today)
2. **Area Chair:** AI Safety Workshop at ICML (2024)
3. **Co-Organizer:** Communication Across Communities in ML Research and Practice at FAccT (2022)
4. **Co-Organizer:** Science and Engineering of Deep Learning at ICML (2021)
5. **Co-Organizer:** Science Meets Engineering in Deep Learning Workshop at NeurIPS (2019)
6. **Organizer:** Theoretical Advances in Deep Learning at Boğaziçi University (2019)
7. **Reviewer:** FAccT, ICLR, ICML, JMLR, COLT, NeurIPS, TPAMI, SIAM Journal; book chapter at Cambridge University Press; grant proposal at NSERC
8. **DEI & Community Work:** Led/participated in initiatives advancing diversity in AI research culture at FAIR, including responsible AI reading groups, inclusive hiring practices, structured mentorship of underrepresented researchers, and cross-community bridge-building events such as with Queer in AI.

Selected Talks & Presentations

AI Scientist and (in-)capabilities of LLMs. Responsible AI Mixer at Microsoft Research NYC (2025)

On generated vs collected data. Global AI Cultures Workshop ICLR (2024)

AI alignment: What is it? Whats new about it? Should you work on it? FAIR (2023)

Can science for ML be value-neutral? Towards a theory of artificial and biological NNs, Les Houches (2023)

On Over-parameterization & Deep Learning Theory (2017 - 2020)

MILA (2019), NYU Shanghai (2019), University of Basel (2019), FAIR New York (2019), Google X (2019), KITP Santa Barbara (2019), Aspen Center for Physics (2019), Google Brain Montreal (2018), FAIR Montreal (2018), MILA (2018), Institut d'Études Scientifiques de Cargèse (2018), ICLR (2018), Heidelberg Institute for Theoretical Physics (2018), Google Brain Zurich (2018), Télécom ParisTech (2017), Université Paris-Sud (2017), FAIR Paris (2017), ENS Paris (2017)

On Optimization dynamics in Deep Learning & High-dimensional Landscapes (2014 - 2018)

NeurIPS Workshop on Integration of Deep Learning Theories Montreal (2018), Poster presentation at ICLR (2018), CILVR Lab Meetings NYU (2018), Information and Computation Seminar NYU (2018), CILVR Lab Meetings NYU (2016), Spotlight talk at Optimization Workshop ICML (2016), MLSS Kyoto (2015), Poster presentation at DLSS University of Montreal (2015), Poster presentation at ICLR San Diego (2015), ML Seminar at Boğaziçi University (2015)

Teaching Experience

Instructor: EM Normandie, Paris. *"What is AI? What can it do?"* (2022)

Invited Lectures: Summer Workshop on Statistical Physics and Machine Learning at Les Houches (2020)

Teaching Assistant: New York University (2011 - 2017)

Courses at the Center for Data Science: *Statistical and Mathematical Methods, Machine Learning.*

Courses at the Courant Institute: *Theory of Probability* (×2), *Probability and Statistics*, *Introduction to Mathematical Analysis II*, *Written Exam Workshop*

Teaching Assistant: Boğaziçi University. *Statistics and Probability* (2009 - 2010)

Additional Training & Summer Schools

Deep Learning & AI Theory (2014 - 2019)

Visiting Scholar at Kavli Institute for Theoretical Physics, UCSB (2019); Statistical Physics and Machine Learning back together, Institut d'Études Scientifiques de Cargèse (2018); Deep Learning and Statistical Physics, Beg Rohu Summer School (2018); Machine Learning Summer School, University of Kyoto (2015); Deep Learning Summer School, University of Montréal (2015); CIFAR NCAP Summer School, University of Toronto (2014); Workshop on Stochastic Gradient Methods, IPAM (2014)

Mathematical Physics Foundations (2009 - 2012)

St. Petersburg School in Probability and Statistical Physics, Chebyshev Laboratory (2012); Minicourse on Compressed Sensing by Emmanuel Candès, University of Cambridge (2011); Summer Schools on Complex Systems (2010), Inverse Problems (2010), and Dynamical Systems (2009)

Fellowships & Awards

MacCracken Fellowship, Courant Institute, New York University (2011 - 2016)
Competitive multi-year doctoral funding
Dora Aksoy Award, Department of Mathematics, Boğaziçi University (2011)
Departmental distinction for top graduating students
Hilmi Tolun Award, Faculty of Arts and Sciences, Boğaziçi University (2011)
Faculty-level distinction for graduating student ranked first in the Faculty of Arts and Sciences
TÜBİTAK Scholarship (2006 - 2011)
Competitive national scholarship supporting undergraduate studies in basic and social sciences
Alize-Barış Tansever Scholarship, Boğaziçi University (2009)
Highly selective award supporting students participating in international exchange programs

References

Research

Max Nickel - `maxn@meta.com`
Research Director at FAIR, Meta
Collaborator and manager

Adina Williams - `adinawilliams@meta.com`
Senior Research Scientist at FAIR, Meta
Collaborator

Giulio Biroli - `giulio.biroli@ens.fr`
Professor of Theoretical Physics at ENS Paris
Postdoc advisor

Léon Bottou - `leonb@meta.com`
Senior Research Director at FAIR, Meta
PhD thesis co-advisor

Gérard Ben Arous - `benarous@cims.nyu.edu`
Professor of Mathematics at CIMS, NYU
PhD thesis advisor

Teaching

Carlos Fernandez-Granda - `cfgranda@cims.nyu.edu`
Associate Professor of Mathematics and Data Science at CIMS & CDS, NYU