Report on the INI Satellite programme "Heavy Tails in Machine Learning" (TML) at The Alan Turing Insitute (2–26 April 2024)

Organisers: O. Deniz Akyildiz (Imperial College London), Anita Behme (Technische Universität Dresden), Emmy Bocaege (The Alan Turing Institute), Emilie Chouzenoux (Université Paris Saclay) Jorge González-Cázares (Universidad Nacional Autónoma de México), Aleksandar Mijatović (University of Warwick & The Alan Turing Institute)

Background and Motivation

The primary aim of the "Heavy Tails in Machine Learning" one-month INI satellite programme at The Alan Turing Institute was to address the emerging phenomenon of heavy-tailed distributions in machine learning models. These distributions, which frequently occur in practical problems involving large datasets and complex models, have a significant impact on model performance, optimisation dynamics, and generalisation capabilities. While heavy-tailed distributions have been observed in various machine learning contexts, a lack of theoretical foundations to explain their emergence and impact posed a significant challenge. The programme aimed to bridge this gap by bringing together experts from statistics, mathematics, and computer science to foster collaborations and drive forward our understanding of heavy-tailed phenomena and develop tools to harness their properties.

Programme Timeliness, Scope, and Outline

The programme was organised at a perfect time and place, given the surge in the use of deep learning and large language models and the need to understand their behaviour in the presence of heavy-tailed distributions. The programme covered a broad range of topics, including:

- Mathematical foundations of heavy tails: Exploring the theory behind heavy-tailed distributions and their impact on stochastic processes.
- Empirical observations in machine learning: Analysing the occurrence of heavy-tailed distributions in machine learning applications, such as stochastic gradient methods and cluster distributions in classification problems.
- Applications in optimisation and neural networks: Investigating how heavy tails affect the convergence of optimisation algorithms and the compression capabilities of neural networks.

The programme consisted of two workshops and an Open for Business event. The opening workshop (TMLW01) focused on the fundamental aspects of heavy tails, their occurrence in different probabilistic settings, and their implications for various machine learning algorithms such as searching for flatter local

minima, obtain better generalisation error bounds and observe compressibility in large neural networks. The closing workshop (TMLW02) delved into the practical applications of heavy tails in deep learning and their role in differential privacy. The Open for Business (OFBW65), which included participants from Google DeepMind, J.P. Morgan and Yale New Haven Hospital, provided a platform for discussing real-world examples of heavy tails in machine learning and a discussion between the current lack of interaction between academia and practitioners working on the same topics. In the appendix below, we provides more details about the activities that took place during the four weeks of the programme.

Scientific Outcomes

The programme yielded several important scientific outcomes:

- Highlights: The programme advanced our understanding of heavy-tailed distributions and their impact on machine learning algorithms (for more details see research outputs and work in progress discussed in the Publications section below). It explored how these distributions emerge, affect convergence rates, and influence model performance.
- Collaborations: The initiative successfully brought together diverse communities, fostering collaborations that resulted in novel insights and potential solutions to longstanding problems. Working groups formed during the programme focused on:
  - Truncation of heavy-tailed random walks: This group investigated how clipping heavy-tailed random variables influences the behaviour of stochastic gradient descent (SGD) algorithms.
  - Compressibility of neural networks: This group explored the connection between heavy tails and the speed of convergence in compressing neural networks.
  - Connection between SGD and SDEs: This group focused on understanding the relationship between SGD and Stochastic Differential Equations (SDEs) driven by heavy-tailed noise and how this influences the dynamics of learning.
  - Heavy tails arising in importance sampling: This group explored the role of heavy tails in importance sampling techniques for rare events.
  - Scale-free random graphs arising in large and deep neural networks: This group investigated the emergence of scale-free random graphs in large and deep neural networks and how this relates to the heavy-tailed nature of weight matrices.
  - ADAM as an alternative to SGD: This group explored the circumstances under which the Adam optimiser, known for its effectiveness in deep learning, outperforms SGD, specially in classification problems in which the class frequency is heavy tailed.
  - Speed of convergence of Levy processes to stable processes: This group examined the speed of convergence of Lévy processes to stable distributions used in machine learning models, focusing on the impact of slowly varying functions.
  - Random matrices and spectrum of neural networks: This group explored the relationship between the spectrum of weight matrices in neural networks and the heavy-tailed distributions that arise during training.
- Advancement in research: The programme contributed significantly to advancing research in the field of heavy tails in machine learning and, particularly, in how to harness the presence of heavy tails.

- Collaboration and community building: The initiative successfully brought together researchers from various fields, from academics to practitioners, fostering collaborations that led to new insights and publications. This was highlighted by the significant publications and working groups that emerged from the programme and continue working together.
- Future Directions: The programme identified several promising avenues for future investigations, including:
  - Generalisation Error in Deep Learning: Understanding how heavy-tailed distributions affect the generalisation capabilities of machine learning models.
  - Robustness and Stability: Investigating the impact of heavy tails on the robustness and stability of machine learning systems.
  - Optimisation Algorithms: Developing new algorithms tailored for handling heavy-tailed data and improving convergence rates.
- Impact beyond academia: The programme generated interest beyond traditional academic circles, influencing areas such as finance, medicine and telecommunications, where heavy-tailed distributions are prevalent.

Publications and/or Grant Applications

The programme has resulted in several noteworthy publications, including:

- Brešar, M., Mijatović, A., Subexponential lower bounds for f-ergodic Markov processes. *Probab. Theory Relat. Fields* (2024). This paper provides new theoretical insights into the convergence of Markov processes in the presence of heavy tails.
- Brešar, M. and Mijatović, A., Non-asymptotic bounds for forward processes in denoising diffusions: Ornstein-Uhlenbeck is hard to beat, arXiv e-prints, 2024. doi:10.48550/arXiv.2408.13799. This preprint provides a tool for comparing denoisiong diffusion probabilistic models in the presence of irregular inital data distributions.

In addition, several collaborations have yielded works in progress, including:

- Propagation of chaos: This collaboration between Aleksandar Mijatović, Mateusz Majka and Umut Şimşekli focuses on understanding the propagation of chaos in machine learning models in the presence of heavy-tailed distributions.
- Compression of neural networks: This collaboration between Jorge Gonzalez Cázares, Bert Zwart, Umut Şimşekli, Kirstin Strokorb and Xingyu Wang aims to optimise compression techniques for neural networks in the presence of heavy tails by obtaining non-asymptotic bounds on these compressions.
- Couplings of heavy-tailed Lévy processes: This collaboration between Jorge González Cázares, David Kramer–Bang and Aleksandar Mijatović investigates the slow convergence phenomenon in the presence of a slowly varying function that is not asymptotically constant and its implications for understanding models that are asymptotically stable and hence heavy-tailed.
- Using rough path theory as a dimensionality reduction technique: This collaboration between Shinpei Nakamura and Hao Ni aims to provide practitioners dealing with high-dimensional time-series with a better alternative to reduce dimensionality that does not incur in the same problems that principal component analysis does.

The programme has also laid the groundwork for several grant applications, aiming to secure funding for continued research in the field. One such example is a PAPIIT grant application made by Jorge González Cázares in UNAM, to obtain computational resources that will enable his team to further pursue research in these directions.

Overall, the programme has made a significant contribution to the field by addressing a crucial problem in the context of machine learning and demonstrating the potential for collaborative research across diverse disciplines.

Appendix

This appendix provides a brief description of the activities during the INI Satellite programme at The Alan Turing Institute.

Week 1 (2-5 April): Courses

We had four courses on topics in Applied Probability and Machine Learning related to Heavy Tails. Video recordings of the lectures and lecture notes for the courses are available on the INI web page.

Prof. Anita Behme (Technische Universität Dresden): Stability of continuous-time processes with jumps. We reviewed the literature on stability of continuous-time processes with jumps: with motivating examples, models and applications, as well as the prerequisites for their analysis. We saw that most solutions to stochastic differential equations are Lévy-type processes and, under certain assumptions, even special Feller. We introduced invariant measures and distributions and reviewed results on infinitesimal invariance, recurrence and ergodicity for Lévy-type processes.

Dr Jorge González-Cázares (Universidad Nacional Autónoma de México): Heavy tails, regular variation and slow convergence. We reviewed the topic of regular variation in the sense of Karamata. Along with the classical Karamata theory, we also discussed its relevance in probability, heavy-tailed distributions, large deviations and and the single-large-jump phenomenon. We then reviewed the literature of stable domains of attraction and showed that, if a Lévy process in the domain of attraction of a stable process possesses a regularly varying component with a nontrivial slowly varying function, then the convergence will be typically slow and this convergence can be arbitrarily slow.

Dr Chang-Han Rhee (Northwestern University): Heavy-tailed large deviations approach to characterising the global dynamics of SGD. We introduced the light-tailed vs heavy-tailed large deviations formulation (conspiracy vs catastrophe principle) as well as its effects on sample-path large deviations. We further saw the fundamental limitations of the classical large deviations principle framework in heavy-tailed systems, introduced M-convergence and the Skorokhod topologies. We presented the heavy-tailed framework for exit-time and metastability analysis (asymptotic atom, uniform M-convergence), characterization of SGD's global dynamics, capitalising on the heavy-tail phenomena in training deep neural networks and the connection between heavy-tails and the edge of stability (multiplicative dynamics and emergence of heavy tails).

Prof. Umut Şimşekli (École Normale Supérieure and INRIA): Heavy-tailed phenomenon in SGD. We introduced statistical learning theory, SGD and neural networks. We reviewed the emergence of heavy-tails in SGD via multiplicative (but possibly light-tailed) noise. We introduced the discretisation of an SDE driven by a stable processes to model SGD. We introduced basic geometric measure theory and fractal geometry and used the Hausdorff dimension of Langevin SDEs driven by stable

processes to obtain generalization bounds. We also saw how neural networks can become compressible if trained with injected heavy-tailed noise.

Weeks 2 & 3 (8–19 Arpil)

Talks

During weeks two and three of the programme we had informal talks, held inside the Ada Lovelace room of the ATI, by the participants on the topics discussed in the working groups listed in Section below.

Bert Zwart: "Robust Mean Estimation" (9 April). *Abstract*: The best way to estimate the expected value of Gaussian data is by using the sample mean. This is no longer true if your data are heavy-tailed. I will explain informally two estimators that are well-known in the literature: the median-of-means estimator, Catoni's M-estimator, and may close with some of my ongoing work in this direction.
References:

- Mean estimation and regression under heavy-tailed distributions – a survey. Gabor Lugosi, Shahar Mendelson, `https://arxiv.org/abs/1906.04280`
- Optimal Mean Estimation without a Variance. Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter L. Bartlett, Michael I. Jordan, `https://arxiv.org/abs/2011.12433`
- Catoni-style Confidence Sequences under Infinite Variance. Sujay Bhatt, Guanhua Fang, Ping Li, Gennady Samorodnitsky, `https://arxiv.org/abs/2208.03185`
- From Data to Decisions: Distributionally Robust Optimization is Optimal. Bart P.G. Van Parys, Peyman Mohajerin Esfahani, Daniel Kuhn, `https://arxiv.org/abs/1704.04118`

Sergey Foss: "Who is responsible for the supremum to be large?" (10 April). *Abstract*: Heavy tails are often associated with the single-large-jump phenomenon or the "catastrophe" principle in which one random variable is responsible for large deviations, as opposed to the light-tailed "conspiracy principle" in which large observations are a result of most random variables exhibiting slightly unexpected behaviours. In this talk, we will see what lies in between: multiple (possibly a non-integer fraction) responsible variables or even a subordinator conditioned on ending at a level after a given time horizon!

Kirstin Srokorb: "A few insights on graphical models in the context of heavy tails" (12 April). *Abstract*: Network structures are at the heart of many data structures, and within the context of this programme even at the data analysis tools themselves. Notably, on the ATI table next to the kitchen there lies a paper "Zoo guide to network embedding". As terminology across different scientific communities varies, some may prefer to speak about graphs instead of networks (I don't mind). In the spirit of the latest presentations, I'd be happy to share a few things I have learned recently about graphical models (from Lauritzen's book), and in particular in the context of, well, should I say "heavy tails"?
References:

- Lauritzen's book: `https://shorturl.at/FT014`
- RSS Discussion Paper: `https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12355`
- New contributions of my own: `https://arxiv.org/abs/2211.15769`

Frederik Kunstner: "Comapring GD, SGD, Adam and Signed descent – why does it work?" (16 April). *Abstract*: We will introduce Adam and its variants and try to give an intuitive account of why and when it works and how does it outperform SGD and its variants. We will also try to explain away some confounding variables such as noise or heavy-tailed transition laws.

Aleksandar Mijatović: "Subexponential lower bounds for f-ergodic Markov processes" (17 April). *Abstract*: We will explain how heavy tails may arise in the stationary distribution of a continuous-time Markov process by establishing lower bounds on its tail probabilities. We will also establish lower bounds on the convergence speed of the Markov process to its stationary law. The techniques are widely applicable and based on Lyapunov functions and sub-/supermartingale methods.

Xingyu Wang: "Importance Sampling Strategy for Heavy-Tailed Systems with Catastrophe Principle" (18 April). *Abstract*: Large deviations theory has a long history of providing powerful machinery for designing efficient rare-event simulation techniques. However, traditional large deviations theory fails to provide useful bounds in heavy-tailed contexts, and designing efficient rare-event simulation algorithms for heavy-tailed systems has been considered challenging. Recent developments in the theory of heavy-tailed large deviations enable designing a strongly efficient importance sampling scheme that is universally applicable to a wide range of rare events. In this talk, we provide an overview of the recent developments in the large deviations theory for heavy-tailed stochastic processes, which is followed by a detailed account of the design principle behind the strongly efficient importance sampling scheme for such processes. The implementations of the general principle are demonstrated through examples in heavy-tailed SGD and settings where the entire path of the underlying system is computationally infeasible to simulate.

Working groups in the "Heavy Tails in Machine Learning" programme

Truncation of heavy-tailed random walks. Team: Bert Zwart, Gerónimo Uribe Bravo, Ceren Vardar Acar, Sergey Foss, Saul Jacka and Bikramjit Das. The following problem is related to clipped SGD. Given iid heavy tailed random variables $X_1, X_2, \dots$ with regularly varying tails of index $\alpha \in (0, 2)$ and $y_n > 0$, it is known that $S_n := \sum_{i=1}^n \min\{X_i, y_n\}$ has tails that can be well approximated by Cramér's theorem when $y_n \equiv y_1$ and have tails of order α when $\liminf_{n\to\infty} y_n/n > 0$. However, it remains to characterise the asymptotic tails of S_n when y_n grows sublinearly. This group of people will also analyse the following 'subproblem'.

Compressibility (speed) of convergence. Team: Jorge González Cázares, Umut Şimşekli, Xingyu Wang and Kirstin Strokorb. Let $X_1, X_2, \dots$ be iid with $\mathbb{E}X_1^2 = \infty$. Define $X^d = (X_1, \dots, X_d)$ and let $X^{d,k}$ be equal to X^d but where all the components that are smaller than the k-th order statistic (in absolute modulus) are removed. It is known that, for fixed $\kappa \in (0, 1]$, we have

$$\lim_{d\to\infty} \frac{\|X^d - X^{d,\lfloor \kappa d \rfloor}\|_2}{\|X^d\|_2} = 0.$$

In this case, we are interested in obtaining convergence rates for this limit and also generalize to when $X_1, X_2, \dots$ are not iid but instead an ergodic Markov process.

Progress has already been made in this direction, establishing non-asymptotic control on the quotient in terms of some expectations that can be empirically estimated with a plug-in Monte Carlo estimator. Further discussions between Umut and others have led to speculate that such results can be more widely applicable within the context of neural networks with heavy-tailed stationary laws.

Connection between SGD and SDG. Team: Ester Mariucci, Mateusz Majka, Gerónimo Uribe Bravo, Ceren Vardar Acar, Saul Jacka, Jorge González Cázares and Aleksandar Mijatović. Following Umut Simsekli's motivation, we are interested in formalising exacly what kind of stochastic dynamics can arise asymptotically as the learning rate in a heavy-tailed SGD η tends to 0. Additionally, they will analyse the possibility that the stationary law of

$$\mathrm{d}X_t = -A_t X_t + \sigma_t \mathrm{d}L_t,$$

is heavy tailed even when (A, σ, L) are light-tailed. Moreover, they will explore if, given some specific heavy tailed limit, specify some processes (A, σ, L) for which that law is stationary via reverse-engineering.

Using some earlier work of Gerónimo on the inverse generator of a one-dimensional stable process, they have come to some asymptotic conditions that will ensure that the stationary distribution satisfies certain tail properties. They further speculate that such an inverse can be computed in d-dimensions by extending some techniques in fractional calculus.

Heavy tails arising in importance sampling. Team: Victor Elvira, Bert Zwart, Mareike Hasenpflug, Sergey Foss. Heavy tails arise naturally in importance sampling problems of rare events, particularly in the Radon–Nikodym derivative. Recent empirical evidence suggests that even truncating the Radon–Nikodym derivative, which introduces a bias, performs well as the bias appears to be controllable. This phenomenon (and other related questions) will be addressed by this team.

Scale-free random graphs arising in large and deep neural networks. Team: William Fitzgerald, Kirstin Strokorb, Gerónimo Uribe Bravo, Ceren Vardar Acar, Sergey Foss. Inspired by recent works by Martin–Mahoney, this team will try to establish rigurously what has been observed empirically. That is, they will try to prove that, under appropriate assumptions on the random high-dimensional matrices arising in large and deep neural networks, the resulting random graph is indeed scale-free. Moreover, they will aim to understand when the eigenvalues, plotted by decreasing magnitude, approximately follow a power-law 'distribution'.

ADAM as an alternative to SGD. Team: Frederik Kunstner, Mareike Hasenpflug and Xingyu Wang. This team will examine when, how and why ADAM outperforms SGD as a stochastic optimiser. We note here that ADAM was inspired by "wrong" ideas (namely, using a second order Newton-type approximation, where the Hessian is thought to be approximately equal to $\sqrt{\mathrm{diag}(\nabla f \nabla f^\top)}$), but actually performs well under certain circumstances.

Speed of convergence of Lévy processes to stable processes. Team: Jorge González Cázares and Aleksandar Miajtović. They are working on obtaining sharp lower and upper bounds on the speed of convergence of Lévy processes in the domain of attraction of a stable process to their limit. In particular, they show that this convergence is typically slow whenever there is a nontrivial slowly varying function that does not have a positive finite limit.

Random matrices and spectrum of neural networks. Team: William Fitzgerald, Frederik Kunstner. Martin and Mahoney [3] observed empirically that the singular values of weight matrices of trained neural networks (for a variety of different architectures) follow a power law and that the exponent in this power law is correlated with the test performance of the neural network. There has been recent interest in relating this observation to the appearance of heavy tails in the limit of stochastic recursions [5], as discussed in the short course delivered by Umut Şimşekli. The empirical observations of [3] motivate two new questions going beyond [5]. What is the behaviour of the spectrum of a matrix

that is evolving under a stochastic recursion equation? How does this depend on the dimension of the matrix, in particular in the limit as this dimension grows to infinity?

It has been found that ADAM generally outperforms stochastic gradient descent (SGD) in training deep neural networks with modern architectures. The theoretical explanation for why this is the case is limited and one-dimensional intuition fails to capture the success of ADAM. A toy model in which the difference between these optimisation algorithms can be explored is to bypass the statistical learning problem and simply to consider a loss landscape given by a 'typical' high-dimensional function. In this setting, typical corresponds to a random function such as a spin glass that have been previously compared successfully to the loss surface of deep neural networks [2]. Comparing the performance of ADAM or SGD on random loss surfaces may give some understanding of which properties of high-dimensional space affect their performance.

In a toy problem of classifying Gaussian mixtures, it has been proved that the trajectory of SGD aligns with the outlier eigenspaces of the Hessian [1]. This builds on numerical work on neural networks [4]. The study of outlier eigenvalues has a long history in random matrix theory starting with the Baik-Ben Arous-Peché (BBP) phase transition. In modern deep learning applications (eg. Chat GPT) the number of classes that are being classified can be as high as 50,000 and the number of observations from each class are imbalanced, often following a power law. This motivates the question of generalising the understanding of outlier eigenvalues within random matrix theory to this dynamical and heavy tailed context.

References

[1] G. Ben Arous, R. Gheissari, J. Huang and A. Jagannath. High-dimensional SGD aligns with emerging outer eigenspaces. Available at arXiv: 2310.03010.

[2] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous and Y. LeCun. The Loss Surfaces of Multilayer Networks. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 38: 192 – 204, 2015.

[3] C. Martin and M. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22, 165, 1–73, 2021.

[4] V. Papyan. Measurements of Three-Level Hierarchical Structure in the Outliers in the Spectrum of Deepnet Hessians. *Proceedings of the 35th International Conference on Machine Learning*, 97:5012–5021, 2019.

[5] Umut Şimşekli, Levent Sagun and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 97: 5827–5837, 2019.

Walk in Hampstead Heath

On Saturday the 13th of April we took a walk in Hampstead Heath, an ancient heath in north London of great historic importance for the city. The park is hilly with ancient and recent woodlands, bathing ponds and the highest point in London.

We met in the morning around 10am and walked from Belsize Park tube station, just south of the park, through the Heath all the way to Kenwood house, close to its northern edge.

This picture of some of the participants of the walk was taken inside Kenwood estate, surrounding Kenwood House visible in the background. Kenwood Hosue, run by English Heritage, is one of London's hidden gems with a stunning world-class art collection and tranquil landscaped gardens.

After lunch in Kenwood Hosue, we enjoyed the permanent collection of classical oil paintings, exhibited in Kenwood House, including works by the masters such as Vermeer and Rembrandt. It was a very relaxing day for the participants, who enjoyed immensly this perhaps less well known London attraction.

Week 4 (22–26 April): Workshop entitled *SGD: stability, momentum acceleration and heavy tails* (22–25 April)

We had 19 speakers (approximately 5 per day) who gave talks on the theme of the workshop in addition to the progress reports. The talks were interesting and very well received. One of the highlights of the schedule was the session devoted to the progress made during weeks 2 and 3 of the programme, where the participants Gerónimo Uribe Bravo, Will FitzGerald, Umut Şimşekli and Xingyu Wang spoke about the progress made by some of the working groups in Section above.

The final day of the workshop (Friday 26 April) was the Newton Gateway event *Connecting Heavy Tails and Differential Privacy in Machine Learning*, that took place in the Wellcome Collection, London, not far from the British Library where The Turing Institute is situated. During this event, we had 4 speakers, 2 from academia (INRIA and Yale) and 2 from the industry (JP Morgan Chase and Google DeepMind) as well as 2 moderated discussions. The talks were well received and participants were highly engaged and active during the moderated discussions.