Generation of heavy tailed distributions with applications to rain generation.

1 Context

State-of-the-art generative models demonstrate remarkable performance on challenging high-dimensional data modalities, such as images [DN], audio [KPH+], and video [HSG+22]. Current implementations predominantly utilize Variational AutoEncoders (VAEs, see [MNG17]), generative adversarial networks (GANs, see [GPAM+]) or denoising diffusion generative models (DPM, see [SSK+21]). However, the standard theoretical guarantees for these approaches usually rely on assumptions that do not hold for heavy-tailed distributions [CDS, TY21].

One such heavy-tailed distribution is the rainfall distribution derived from radar imaging of specific regions. While one might expect methods effective for typical image distributions to be well-suited for the task at hand, the unbounded and potentially heavy-tailed nature of the rainfall distribution poses significant challenges. Moreover, current research on climate change suggests that the likelihood of extreme rainfall events is expected to increase in the future. Therefore, generating data from the tail of this distribution is a critical task for urban planning and insurance.

Recently, [LNF23] illustrated that standard VAE are not well suited to explore extreme value. They introduced a new architecture to support that when appropriately tailored and trained, VAE can explore heavy-tailed distributions. However, this approach relies on a particular decomposition of the target distribution between an to univariate extreme variable and a non extreme component and might not be adapted to time and space-dependent data. On the other hand, [SSD24] proposed new diffusion-based samplers using α-stable noise in particular to sample from heavy-tailed target distributions and [AGG22] also proposes a GAN with different latent distribution to sample from heavy-tailed targets. Those works present encouraging empirical results, but no theoretical guarantee is offered for the reconstruction of the tails and the application to rainfall fields is an open question.

The guiding dataset for the evaluation of the proposed approaches will be two open source datasets of rain radar images from both a region in Germany and from the Swiss Alps. While the first can be considered almost stationary (spatial homogeneous), the second is highly non-stationary due to orographic effects. In the context of this internship, we will focus on aggregated daily data, concerning either radar images or a subset of rain stations.

Internship goals

The internship goal is to explore a subset of the following research lines.

- Understand the different theoretical guarantees holding for GAN, VAE and Diffusion models and identify their limitations when handling heavy-tailed data.
- Identify the main limitations of the existing algorithms in the specific benchmark datasets (homogeneous vs non homogeneous, tail vs bulk generation, etc.) and explore new architectures adapted to such data.
- Explore adaptations of current proof strategies in the existing literature to adapt the frameworks to handling heavy tailed data. A natural first step will be to investigate VAE for heavy-tailed distributions.

Practical information

The internship is part of a collaboration between the LPSM (Sorbonne Université), Geostatistics of Mines Paris PSL and BioSP INRAE labs and financed by the Geolearning chair. The internship will take place either in the LPSM or the Geostatistics lab with occasional visits to the BioSP team in Avignon.

Contact:
Sylvain Le Corff (sylvain.le_corff@sorbonne-universite.fr)
Gabriel Victorino Cardoso (gabriel.victorino_cardoso@minesparis.psl.eu)

References

[AGG22] Michaël Allouche, Stéphane Girard, and Emmanuel Gobet. EV-GAN: Simulation of extreme events with ReLU neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022.

[CDS] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions.

[DN] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. 34:8780–8794.

[GPAM+] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680. MIT Press.

[HSG+22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[KPH+] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*.

[LNF23] Nicolas Lafon, Philippe Naveau, and Ronan Fablet. A vae approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*, 2023.

[MNG17] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International conference on machine learning*, pages 2391–2400. PMLR, 2017.

[SSD24] Dario Shariatian, Umut Simsekli, and Alain Durmus. Denoising levy probabilistic models. *arXiv preprint arXiv:2407.18609*, 2024.

[SSK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

[TY21] Rong Tang and Yun Yang. On empirical bayes variational autoencoder: An excess risk bound. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4068–4125. PMLR, 15–19 Aug 2021.