Book review

Synthetic Data for Deep Learning, by Sergey I. Nikolenko, Springer (2021). E-book. ISBN 978-3-030-75177-7.

When Deep Learning was just about entering the domain of Mathematical Finance, it was already evident for many that data would soon assume a more significant role in financial modelling, be it to make predictions, analyse the market or to train models. What was much more difficult to anticipate—at least to the majority of users—was the emerging role of synthetic data in this context. Given the terabytes of data footprints produced on a daily basis in an array of domains, we were (quite unsurprisingly) speaking about big data, rather than synthetic data for a long time. It was only as deep learning applications became more prevalent, that use cases (and thus the need) for high quality synthetic data became important.

Though interest for this area has recently grown through many channels, for many of the people following research in Quantitative Finance it was the work Deep Hedging (Buehler et al. 2019) that kickstarted interest in synthetic data. Deep Hedging was one of the applications that brought the advantages of synthetic data (or as recently coined by the domain ‘Market Generators’) to light. The interest in market generation has opened up new avenues for financial modelling and has led to a surge in research activity in the domain with a number of recent contributions (for example Bonnier et al. 2019, Kondratyev and Schwarz 2020, Kondratyev et al. 2020, Wiese et al. 2020, Acciaio et al. 2021, Buehler et al. 2021) and doubtless more to come. It is thus no surprise that several of the 2022 Risk Awards went to Hans Bühler and the JP Morgan team. In more traditional Machine Learning applications (think robotics or autonomous driving) the case for synthetic (training) data was made far earlier than in finance. Nevertheless, even in the more classical domains, only fairly recent developments have focussed attention (or an entire book) to the study of synthetic data for deep learning: In fact, the author joyfully exclaims in the preface that he ‘managed to release one of the first books specifically devoted to the subject of synthetic data’. And while Nikolenko goes on to convince the reader why this subject deserves our attention in general, we devote this book review to discussing why this subject deserves
the attention of quants and quantitative finance researchers, specifically.

_Synthetic Data for Deep Learning_ by Sergey I. Nikolenko appeared in 2019 in the Springer series _Springer Optimization and Its Applications_. The book is slightly outside of the range of our regular themes and it is in fact strictly speaking not a book for Quantitative Finance: To put it differently, it is not a book that was written with the aim of addressing questions arising in finance. It is a book on deep learning in its more traditional sense and the applications are more aligned with typically known use cases of deep learning: from computer vision problems to optical flow estimation and navigation, none of which are related to finance in any obvious form. So why then was it selected to be in the spotlight in this journal, especially if several of the topics discussed in the book are not (yet) on the standard agenda of its community? It is a distinctive flavour of quantitative finance that one can—and perhaps indeed should—get inspiration by doing some window-shopping in other disciplines and immersing oneself with new methods on display there. Such excursions can (and in the past quite frequently did) add to the quantitative modeller’s tool-kit. It may well be that the book will provide some inspiration for useful new tools in this spirit. With this in mind, we consider _Synthetic Data for Deep Learning_ a great read for those who are currently interested or active in developing synthetic data solutions in quantitative finance, and in particular to those who do not mind being left hungry for answers to questions that arise along the way.

The author, Sergey I. Nikolenko is beyond doubt very active in the area with a previously published monograph, which shares its title ‘Deep Learning’ with the famous 2016 reference work (Goodfellow et al. 2016) by Goodfellow, Bengio, and Courville. Nikolenko is Head of Lab at the Steklov Institute of Mathematics at St. Petersburg, and has been commercially active for a number of companies, amongst them large international businesses as well as smaller providers of Machine Learning solutions.

Praise and criticism: This book has some potential to inspire new tools for finance since it showcases several applications and an array of challenges where synthetic data is used for deep learning. It also displays some of the typical pitfalls in those applications. However if the book does inspire new tools, they are quite certainly not ready-made for quantitative finance and will require some adaptation by the prepared reader. It should also be pointed out, as the does author himself, that this is not an introductory textbook. Although it contains several introductory chapters on deep neural networks and corresponding optimisation problems and some on deep generative models as well as neural architectures for computer vision, these introductory chapters are better seen as a reminder rather than a thorough introduction. The book is targeted to ‘a somewhat prepared reader’ who will only use the introductory chapters as reference material. It is disappointing to this reviewer that the author, who is clearly practically oriented and experienced, has not shared more practical examples and in particular, code. It would have been immensely useful to have snippets of code available alongside the chapters, or pointers to repositories which would have helped the interested reader to actively engage with the concepts.

Structure and contents: The book itself is structured into ten chapters in addition to an introductory and a concluding chapter. The introductory chapter picks up what the author promises to do: It explains ‘the data problem’ to convince readers that synthetic data deserves their attention in the first place. The concluding chapter (chapter 12) points to possible directions for future work and to domain adaptation. The contents cover three main directions for the use of synthetic data in machine learning and provide a high-level walkthrough of typical challenges and solutions (albeit from a perspective unrelated to finance):

1. Using synthetically generated datasets to train machine learning models directly.
2. Using synthetic data to augment existing real datasets so that the resulting hybrid datasets are better suited for training the models.
3. Using synthetic data to resolve privacy issues that make the use of real data difficult.

The discussions around synthetic data highlight a question relevant for finance from a regulatory perspective i.e. whether realism in synthetic data research is always as necessary as we tend to think, in order to efficiently train models. The book revisits this question several times with different viewpoints in different applications (though also here answers in the financial context are yet to be delivered). It is somewhat inconvenient for the curious reader that these themes are mainly presented from a computer vision angle. Although the book touches on financial aspects (Chapter 11) in a rudimentary fashion in the context of privacy issues, point (3) above. In this chapter, there are a number of blanks left to the finance-inclined reader’s imagination. However, the reality is that topics around synthetic data and questions of data privacy in finance are currently picking up and gaining momentum among quants and quantitative finance researchers and there is a lot to say in this area: We can expect to see many more finance-related research in this area in the future.

What the reader can expect to gain from this book, will strongly depend on their previous knowledge. The author elegantly acknowledges this by distinguishing between ‘unprepared’ and ‘prepared’ readers. As pointed out previously, it is not ideal as an introductory textbook: Often the fuzzy details and tedious discussions are circumvented for the sake of brevity and a neat presentation. On the plus side, readers may find themselves getting high-level insights whilst skimming through chapters. However, the same readers (either prepared or unprepared) may find it somewhat difficult at times to fill in the details or to reproduce the numerical results independently. Nevertheless, all readers of this book will gain a broader perspective on generative modelling and its historical evolution. Perhaps even more importantly, this book will give its readers a good starting point to communicate with and build bridges across disciplines through an increased awareness of the terminologies used in other areas of applications and the typical challenges and pitfalls that arise. We are convinced that such awareness will bear numerous benefits in an ever more interdisciplinary research landscape.
References

Acciaio, B., Munn, M., Wenliang, L.K. and Xu, T., COT-GAN: Generating sequential data via causal optimal transport. Adv. Neural Inf. Process. Syst., 2021, 33, 8798–8809.

Bonnier, P., Kidger, P., Perez Arribas, I., Salvi, C. and Lyons, T., Deep signature transforms. Adv. Neural Inf. Process. Syst., 2019, 32, 3082–3092.

Buehler, H., Gonon, L., Teichmann, J. and Wood, B., Deep hedging. Quant. Finance, 2019, 19(8), 1271–1291.

Buehler, H., Horvath, B., Lyons, T., Perez, I. and Wood, B., Generating financial markets with signatures. Risk Magazine, June Issue, 2021.

Goodfellow, I., Bengio, Y. and Courville, A., Deep Learning, Adaptive Computation and Machine Learning Series, 2016 (MIT Press: Cambridge, MA).

Kondratyev, A. and Schwarz, C., The market generator. Risk Magazine, February Issue, 2020.

Kondratyev, A., Schwarz, C. and Horvath, B., The data anonymiser. Risk Magazine, August Issue, 2020.

Wiese, M., Knobloch, R., Korn, R. and Kretschmer, P., Quant GANs: Deep generation of financial time series. Quant. Finance, 2020, 20(9), 1419–1440.

Blanka Horvath
Technical University of Munich
© 2022, Blanka Horvath

Blanka Horvath, Prof. Dr, is Assistant Professor for Financial Mathematics at the Technical University of Munich and core member of the Munich Data Science Institute, leading the SyBenDaFin project which drives efforts towards Synthetic Benchmark Datasets for Finance. Since 2019 she has been active as a visiting researcher at The Alan Turing Institute building the Finance and Economics programme’s Machine Learning in Finance theme and co-organising the programme Synthetic data generation for finance and economics.