Stephan Mandt **Research Statement** **July 12, 2021**

Most recent advances in artificial intelligence have been based on deep neural networks trained on massive datasets, and there is an ever-growing demand for bigger and more powerful models. However, increasing the power of Deep Learning (DL) entails growing energy and storage costs due to massive data needs and parameter-rich models. For instance, recent models for natural language generation contain more than a hundred billion parameters, so that their training requires nearly five times the lifetime carbon dioxide emissions of the average car. The main goal of my research is to **solve resource inefficiencies in a wide range of contexts** through a holistic approach: I view DL models and algorithms through the lens of information theory and Bayesian inference, whereby I both quantify and minimize the required resources.

My research primarily focuses on **deep latent variable models** and their training paradigms, as well as on **probabilistic DL** in general. Within probabilistic DL, my research seeks to improve three types of resource inefficiencies. (1) *Runtime inefficiency*: a Bayesian learning or inference algorithm's inability to produce desired answers within a given computational time budget, (2) *data inefficiency*: a model's excessive reliance on training examples, and (3) *bandwidth inefficiency*: a model's inefficient representation of parameters or data. To promote runtime efficiency, I develop **variational inference** algorithms that are faster than conventional Bayesian approaches. To increase data efficiency, I integrate deep latent variable models with **informative priors** to connect different learning tasks. Finally, to improve bandwidth efficiency, I develop **neural compression algorithms** for both data (images and video) and model parameters.

An additional goal of mine is to make probabilistic DL accessible beyond the computer science community to spur innovation in a wide range of fields. Therefore, I apply probabilistic DL in the **natural sciences**, particularly in physics, chemistry, and climate science.

In what follows, I will first summarize the primary contributions along these lines and provide more detail on my research.

Publications. Overall, I co-authored 32 papers in top-tier machine learning conferences with acceptance rates of 20-25%. These conferences include ICLR, NeurIPS, and ICML, which are considered to be the most influential and competitive publication venues in machine learning[1]. In addition, I co-authored 13 journal papers in machine learning, physics, and chemistry and hold seven patents. My h-index is 24, and my work has been cited 2600 times at the time of writing. The website csrankings.com shows that my research significantly contributes to UCI's ranking in machine learning.

Funding. I have acquired a total amount of around 3 million US dollars in grant funding as a PI or Co-PI over the course of the tenure period, with around **USD 1.85 million** dedicated to my research group. For instance, I am the PI of a large **DARPA** project on novelty detection, where I represent three UCI groups. In addition, I have been awarded two NSF grants as sole PI: the **NSF CAREER Award** and an **NSF RI Small Grant** (acceptance rate ~10%). I am also a co-PI on two other NSF grants on supervised compression and on the future of work, respectively, that have similar competitive acceptance rates. Furthermore, I have acquired continued funding from Qualcomm over three successive years.

Talks. I am frequently invited as a speaker at top-tier events, including two NeurIPS workshops, the National Academy of Sciences' Kavli Frontiers of Science Symposium, and a Dagstuhl seminar; I was also invited as a keynote speaker at a leading German ML conference (LDWA). In addition, I gave invited talks at various universities, such as CMU, Caltech, Vector Institute, and ETH, as well as corporate research labs, such as Google Brain, FAIR, Qualcomm, Adobe, Snap, Intel, and IBM. More details can be found on my CV.

In the next section, I describe my research contributions in more detail, presenting them through the lens of improving bandwidth, data, and runtime inefficiency.

[1] The top three conferences in AI according to https://scholar.google.de/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence

(i) Advancing Variational Inference for Runtime-Efficient DL.
While classical machine learning seeks the optimal parameters of a model, Bayesian machine learning aims to identify *all possible* parameter configurations compatible with an observed dataset. The resulting distribution over compatible parameters is called the (Bayesian) posterior; it can be used for model averaging or uncertainty assessment and can guide quantization and compression. For most models, the posterior is intractable to compute without additional approximations. While Markov Chain Monte Carlo (MCMC) approaches generate samples from the posterior, these algorithms can still suffer from slow convergence. Variational inference (VI) algorithms are a faster alternative to MCMC; they approximate the posterior by efficiently solving an optimization problem with stochastic gradient descent (SGD).

My research has shown connections between VI, MCMC, and SGD algorithms. In "Stochastic Gradient Descent as Approximate Bayesian Inference," I analyzed SGD with constant learning rates ("constant SGD") in a continuous-time limit (W3) and showed that this algorithm simulates from an approximate posterior (C4, J9). Specifically, I showed how to tune the parameters of constant SGD to best approximate the posterior by solving a variational problem. The same continuous-time formalism proved useful to quantify discretization errors and other properties of MCMC algorithms. This work has been cited 450 times to date, and its proposed algorithms have been integrated into Google's Tensorflow Probability framework. When visiting Google Brain, I expanded this line of research further. For example, in "How Good is the Bayes Posterior Really?" (cited 75 times in less than one year), we postulated the "cold posterior effect" that sheds light on the failures of current MCMC approaches in large-scale DL (C25).

Overall, making VI approximations more practical and confident has been among the main themes of my research, and I have written a well-cited review article on the topic (J10). For example, we improved the posterior approximation obtained in “amortized” VI significantly by adopting an iterative meta-learning framework (C15). My research has furthermore integrated diverse ideas such as Quasi-Monte Carlo sampling (C16), low-rank approximations (C24), tempering (C3), physics-inspired perturbation theory (C9, J11), and variance-reduced optimization (C1) into VI. I have also played a lead role in the annual symposium on advances in approximate Bayesian inference since its foundation in 2014.

(ii) Deep Probabilistic Modeling for Data-Efficient DL.
Deep neural networks often rely on vast amounts of training data, which can be expensive to collect. DL models that rely on excessive amounts of training data are furthermore often vulnerable to changes in the data distribution (e.g., data from different hospitals or images taken in changing light conditions). Such failure can have dramatic consequences for decision-making, e.g., in medical prediction tasks or autonomous driving, respectively.

Increasing data efficiency in DL can be approached from different angles and may involve imposing structural (e.g., logical, causal, or symmetry) constraints on architectures, actively selecting the most informative data points (C7, C17), or transferring knowledge across domains. My research mostly focuses on the latter perspective, where I construct **informative priors** for deep latent variable models to guide their learning tasks. Examples include augmenting VAEs with time series priors for video forecasting (C13, W18), matrix factorization priors for predicting audience face reactions to movies (C6), scalable Gaussian Process priors (C30), or point process priors for event sequence forecasting (C28).

Another benefit of incorporating informative priors into latent spaces is to help a model find interpretable patterns. E.g., in "Dynamic Word Embeddings" (C8), we hybridized a probabilistic interpretation of word2vec with latent Kalman filters for learning word embeddings over time. By training the model on Google books, we traced and visualized the evolution of word meanings over large periods (e.g., “simulation” moving from “deception” to “computer”). This work has been cited 180 times to date. Our technical accomplishment was a new structured variational smoothing scheme for non-conjugate latent time series models that we improved in (C14). We adopted a similar approach for imputing missing values in time series with Gaussian Process VAEs (C23) and generalizing temporal priors in dynamic topic models (C10).

(iii) Neural Compression for Bandwidth-Efficient DL. Deep neural networks incur immense computational and storage costs. They typically involve a large number of weights and/or high-dimensional feature representations of the data. Both properties make it expensive to communicate features and store the model on resource-constrained devices. As follows, I summarize these inefficiencies related to data communication, processing, or storage cost as *bandwidth inefficiencies*.

To tackle bandwidth inefficiencies with DL, I work on neural compression, where I focus both on data and model compression. Specifically, my research draws on a formal equivalence between the objective that variational inference optimizes (the "ELBO") and the rate-distortion objective in lossy compression. My research on VI for compression has led to state-of-the-art results on both image (C29) and video coding (C31). I was awarded three NSF grants related to this topic, and I am also authoring three patents (P6, P4, P3).

A representative example of how advances in VI can guide lossy compression is “Improving Inference for Neural Image Compression” (C29), which identified three performance gaps: an amortization gap (a shortcoming of amortized variational inference), a discretization gap (due to the mismatch between the training and testing objective), and a marginalization gap (related to the fact that latent variables cannot be tractably integrated out). Our paper provided solutions to narrow each gap, resulting in 15-20% bitrate savings in established benchmark models.

Overall, my compression research covers a variety of projects lying on a spectrum between theoretical and applied. For example, we published one of the first papers on neural video compression (C21), drawing on earlier work on disentangling static from dynamic information in sequential data (C13). On the theoretical end of the spectrum, we explored the best achievable compression performance in trained Bayesian models (C26) and the possibility of bounding the rate-distortion function of a data source based on samples (W22).

(iv) Machine Learning and Science. To promote the impact of probabilistic DL in other fields, I work on applications in the natural sciences. Trained as a physicist with work on random matrix theory (J1) and statistical mechanics (J2-J7), I have expanded the scope of my research to physics, chemistry, and climate science more broadly.

Many problems in the natural sciences involve data sparsity and are therefore suitable for latent variable modeling (see section (ii)). For example, chemical engineers are interested in predicting the properties of fluid mixtures, such as alcohol and water. However, since experiments are expensive, most possible fluid combinations are still unexplored. In J12, we outperformed established physical models using a matrix factorization approach and showed that both models could be hybridized (J13). This project has led to an ongoing collaboration on predictive thermodynamics with a German university.

Other fields show an abundance of data but require tools to discover latent structure. Climate scientists, for example, are interested in identifying low-dimensional statistics that capture how their systems respond to changing conditions (e.g., global temperatures). In C27, we analyzed the effects of geography and temperature on cloud formation and convection, using deep unsupervised models; we are currently generalizing our analysis to study the impact of global warming.

Last but not least, physics can be a source of inspiration for ML. For example, spontaneous symmetry breaking and the Higgs mechanism led us to identify and solve an optimization problem in time series models (C14).

Concluding Remarks.

Deep Learning has achieved enormous performance gains, but the current trend of exponential model growth is not sustainable in a world with limited resources. The need for more resource-efficient DL also goes hand in hand with an increasing demand for small-scale models that fit on low-powered devices, such as drones or cellular phones, as well as new solutions to data compression. In my research, I approach different types of resource inefficiencies in probabilistic DL by drawing on the toolboxes of information theory and approximate Bayesian inference. By addressing the dimensions of runtime, data, and bandwidth efficiency, my research has established new algorithms for faster inference in deep probabilistic models, informative priors in deep latent variable models for structured (e.g., sequential) data, and state-of-the-art neural image and video compression approaches. In addition to computer science applications, I have applied these methods in the natural sciences, in particular physics, climate science, and chemistry. Ultimately, my research contributes to making probabilistic deep learning as efficient and easy to use as non-probabilistic DL while dramatically reducing the required resources.

References

References refer to the numbering system of my CV, also available at http://www.stephanmandt.com/.