How Variational Autoencoders Came To Be:
A Commentary by D Kingma

James V Stone
Email: *j.v.stone@sheffield.ac.uk*
File: KingmaCommentaryVAE2025.tex.

April 25, 2025

Note: This is an addendum to the book: The Artificial Intelligence Papers by JV Stone `https://jamesstone.sites.sheffield.ac.uk/books/theaipapers` .
Reference: Kingma, D. P. and Welling, M. (2013). *Auto-encoding variational Bayes*, which can be viewed at `https://arxiv.org/pdf/1312.6114v11.pdf`.

My initial interest in generative modeling was sparked by the works of Geoffrey Hinton and his students, who wrote influential papers on generative models since the 1990s. Before I started my Ph.D., I used to check his website daily for novel works. Geoff wrote many influential papers on this topic, including papers on the EM algorithm and the lower bound of the likelihood. One key idea is expressed in the title of his 2006 paper "To Recognize Shapes, First Learn to Generate Images". I think this may have, in turn, been inspired by "What I cannot create, I do not understand" which was famously found written on Richard Feynman's blackboard after his death.

The general idea is that in order to fully understand data, the goal of a neural network should be to maximize the joint probability over all dimensions of the data. If a model can do this well, it has learned the underlying structure of the data, which entails that it must have learned useful feature representations of the data. I found this an incredibly beautiful and compelling insight. This can be alternatively formulated as trying to maximally compress the data. This can be viewed as a form of Solomonoff induction, and Occam's razor. It felt right.

One other aspect of Geoff's work is that of latent variables. The Helmholtz Machine from 1995 can be seen as a precursor to variational autoencoders (VAEs), one difference being that the Helmholtz Machine used discrete latent variables, and was not optimized towards maximum likelihood, but wake-sleep.

In my free time I worked on building predictive models for prediction markets. I learned that in order to build a well-functioning model, i.e. in order to get accurate odds, it's important to be very rigorous: the math needs to be exactly right. In addition, a good strategy for the problem I worked on was to maximize the joint probability of the data over all the variables. This further convinced me of the usefulness of generative models, and specifically maximum likelihood.

In 2009, for my Master's thesis, I worked as a junior researcher with Yann LeCun at NYU on neural networks and generative modeling. I felt right at home. (AlexNet came out during my tenure at NYU. Yann called for an emergency meeting with all his staff, since he felt that it should've been his lab, not Geoff Hinton's lab, although they were good friends of course. Completely understandable, since feedforward convolutional nets were much more in Yann's lane than Geoff's lane.) In 2012, in preparation for my PhD, I worked with Yann again. My intuition was that we should train feedforward neural networks (as opposed to RBMs, which were dominant at the time) as generative models with latent variables by maximizing likelihood.

Yann was not sold; in his view, maximum likelihood was just a special case of energy-based models, and we should just treat the output of models as energies, as opposed to log-probabilities. My view, in contrast, was that while it's true that log-probabilities are indeed a special case, it's a special case with very useful special properties. In addition, Yann explained he is "allergic" to sampling-based approaches. At the same time, I discovered that Max Welling was moving to the Netherlands, and starting a lab there, and had an open position for a single Ph.D. student. I applied, and got the position. Max Welling's research interests were greatly aligned with mine.

While in New York, I was also intrigued by dropout, which recently came out, and I was wondering how we could interpret it probabilistically. I figured that we could view dropout's Bernoulli noise as binary latent variables, while at the same time we could view the network's stochastic activations as latent variables, just reparameterized. The objective can then be seen as a Monte Carlo estimator of the (conditional) maximum likelihood.

I started my Ph.D. position in Amsterdam the first half of 2013. I couldn't have wished for a better Ph.D. advisor than Max. We had a weekly meeting where we would discuss new results. Max had high standards, pushed me to be rigorous, expecting me (as he did with all his students) to describe the mathematics and prove results on his whiteboard on the fly. In our weekly meeting, he was able to grasp problems quickly, and give pointed feedback.

I felt that the solution to training latent-variable models should make use of the backpropagation algorithm: for only twice the cost of a forward pass, one can compute full gradient information w.r.t. the inputs and/or parameters of a model. Can we use this to more efficiently model the posterior? For example, we could do Hamiltonian Monte Carlo using backpropagation-provided gradients to sample from the posterior. This is what I initially did.

I did experiments where I investigated how efficiently MCMC-based posterior inference worked with neural networks with latent variables in their original versus reparameterized form (inspired by dropout). In June of 2013, I published a single-author paper on the topic. Later, with Max, we wrote a paper on the topic that we published at ICML.

In 2013 I was helping Max teach a course on machine learning. As part of this course I had to grok Chris Bishop's book, "Pattern Recognition and Machine Learning". From this book I learned some interesting concepts, including variational inference. In variational inference, the objective is the variational lower bound. i.e. the lower bound the marginal likelihood using an expectation over tractable log-probabilities. This estimator of the variational lower bound becomes a close approximation to the marginal likelihood, if we optimize the posterior sufficiently.

It seemed to me that it would be great if we could optimize feedforward neural networks with continuous latent variables using the variational lower bound. And instead of optimizing a separate posterior per datapoint, it would be much more scalable if we could train a neural network to predict the optimal approximate posterior, like Helmholtz machines, but properly using the variational lower bound.

While on my bike on my way to university, the idea came to me that we could just reparameterize the variational lower bound, such that the noise becomes external and we can just backprop through the latent, similar to dropout. The same idea could be used to estimate a posterior over the parameters, not just the latent variables.

I did experiments on it, and it seemed to work! I explained the idea to Max, but thought it might be too simple for a paper. Max countered that it was not. He also suggested that we should publish quickly, since he said that similar ideas might be going around in other labs. We submitted the paper in December of 2013, and uploaded it to arXiv.

I think that, due to their elegance and generality, VAEs are the most beautiful result in my own research career. Ten years later, the paper would win the inaugural ICLR “Test of Time” award. I feel extremely blessed by the amount of impact we made and recognition we have received for the work.

The work opened up a new field of machine learning research, resulting in a plethora of interesting follow-up works, in both machine learning and various practical applications in the sciences, including medicine. In the intervening years, Max and I published various works extending VAEs to hierarchical models and more expressive posteriors. I also later worked on diffusion models, now popular for image, video and audio generation, showing under which conditions continuous-time diffusion models can be viewed as a special case of hierarchical VAEs with an infinite number of layers. I believe there are still plenty of exciting results left to be discovered, especially on how to scale VAEs to the very large-scale regime.

Durk Kingma, April 2025.