# Variational Inference and Deep Generative Models

Wilker Aziz and Philip Schulz

## 1 Description

NLP has seen a surge in neural network models in recent years. These models provide state-of-the-art performance on many supervised tasks. Unsupervised and semi-supervised learning has only been addressed scarcely, however. Deep Generative Models (DGMs) make it possible to integrate neural networks with probabilistic graphical models. Using DGMs one can easily design latent variable models that account for missing observations and thereby enable unsupervised and semi-supervised learning with neural networks. The method of choice for training these models is variational inference.

This tutorial offers a general introduction to variational inference followed by a thorough and example-driven discussion of how to use variational methods for training DGMs. It provides both the mathematical background necessary for deriving the learning algorithms as well as practical implementation guidelines. Moreover, we discuss common pitfalls that one may encounter when using DGMs for NLP applications, such as the latent variable being ignored by the model, and discuss potential solutions from a theoretical and practical perspective. Importantly, the tutorial will cover models with continuous and discrete variables.

The tutorial starts by motivating the need for approximate inference methods from well-known models such as LDA (Blei et al., 2003) and factorial HMMs (Ghahramani and Jordan, 1996). It then derives the variational objective (evidence lower bound) in detail and relates it to expectation maximisation. This is done to provide the audience with a point of reference as we assume that the EM algorithm will be familiar to most participants.

Next, we present the general idea of a DGM and explain why these models have not been widely used until a couple of years ago. We illustrate the problem with a discussion of the wake-sleep algorithm (Hinton et al., 1995). The next part of the tutorial is dedicated to variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014). We derive the Gaussian reparametrisation that makes it possible to sample stochastic gradient estimates and thus use backpropagation for training (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014). We then discuss conditions under which this simple training procedure may fail and offer modifications (e.g. downscaling the KL-term of the variational objective Bowman et al., 2016).

In the last part of the tutorial, we focus on discrete latent variable models which can generally not be trained using a reparametrisation. Instead they require the score-function gradient technique (Williams, 1992; Blei et al., 2012). This is yet another way of sampling a stochastic gradient that works for discrete as well as continuous variables. It does not change the variational objective but does suffer from higher variance than the reparametrisation gradient. We will therefore discuss control variates and baselines as basic variance reduction techniques. Providing the audience with the necessary tools for constructing DGMs with continuous and discrete variable will enable them to conduct research with DGMs without being limited to only one set of distributions.

The slides used in the tutorial can be viewed and downloaded from `https://github.com/philschulz/VITutorial`. The slides for discrete latent variables will be added soon.

The tutorial is complemented by a practical coding exercise. The exercise is implemented in a Python notebook and provides the user with a step-by-step walkthrough for the creation of a variational autoencoder. The notebook can be downloaded from `https://github.com/philschulz/VITutorial/blob/master/code/vae_notebook.ipynb`. While we will not have enough time to do the coding exercise during the tutorial, the audience will be encouraged to do the exercise in their own time and contact us with any questions they may have. Additionally, we also provide short notes on the more intricate mathematical details that the audience can use as a reference after the tutorial. We expect that with these additional materials the tutorial will have a long-lasting impact on the community.

## 2 Outline

| **Time** | **Content** |
|---|---|
| 10 mins | Welcome and motivation for approximate inference techniques |
| 20 mins | Derivation of variational inference and relationship to EM |
| 10 mins | Mean field variational inference |
| 10 mins | Introduction to Deep Generative Models |
| 20 mins | Variational Autoencoder |
| 20 mins | Gaussian reparametrisation and stochastic gradient sampling |
| 15 mins | Problems with training DGMs |
| 15 mins | Coffee break |
| 20 mins | Semi-supervised learning in DGMs and the need for discrete latent variables |
| 20 mins | Score-function gradient (REINFORCE) |
| 10 mins | Examples of models that use the score-function gradient |
| 10 mins | Wrap-up and goodbye |

## 3 Instructors

1. Wilker Aziz

    - Affiliation: University of Amsterdam
    - E-mail: w.aziz@uva.nl
    - Website: http://wilkeraziz.github.io
    I am a research associate at the University of Amsterdam working on natural language processing problems such as machine translation, word alignment, textual entailment, and paraphrasing. My interests sit at the intersection of disciplines such as formal languages, machine learning, approximate inference, global optimisation, and computational linguistics. Recently, I've developed quite an interest in Bayesian deep learning. In particular, I'm developing probabilistic neural network models that reason with and induce forms of discrete generalisation such as trees and graphs.

2. Philip Schulz

    - Affiliation: University of Amsterdam
    - E-mail: P.Schulz@uva.nl
    - Website: philipschulz.org
    I am interested in Bayesian graphical modelling, deep generative models, and approximate inference methods. Currently I am working on hierarchical Bayesian models for machine translation and in particular word alignment. These models require efficient inference algorithms such as MCMC sampling and variational inference. Together with Wilker I have designed a fast Gibbs sampler for our word alignment model. More recently, I have started to work with variational inference and stochastic gradients.

# 4 Estimated Audience Size and Previous Presentations

Parts of the tutorial were given for individual research groups and averaged about 20 people per presentation. Given the current interest in deep generative models and deep learning more generally we would expect an audience size of roughly 100 people.

- Monash University (Melbourne, Australia): November 16 2017
- University of Melbourne (Melbourne, Australia): October 31, November 02 and 07 2017
- Amazon (Berlin, Germany): July 26 and 27 2017

## 5 Techical Requirements

Wifi, a board or flip chart, and whiteboard markers are the only technical aids required.

# 6 Venues

The preferred venue for this tutorial is ACL 2018. The reason is that we expect an audience with a strong technical background and a keen interest in new machine learning techniques at this conference. EMNLP 2018 would also be possible for the same reasons. However, we reckon that we can reach a larger audience and thus better disseminate these new thechniques at ACL. In particular, ACL attracts a more balanced mix of technical and non-technical audiences and thus gives us the chance to relate Deep Generative Models to researchers who may not even be aware of them.

NAACL and Coling are dispreferred options since the former is regional and the latter has a strong focus on linguistic insights. While we do believe that Deep Generative models can enable and enhance linguistic investigation of language data, we are not sure that the audience at Coling would feel attracted to a machine-learning-focused tutorial like ours.

# References

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003. ISSN 1532-4435. doi: 10.1162/jmlr.2003.3.4-5.993. URL http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

David M. Blei, Michael I. Jordan, and John W. Paisley. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012. URL http://icml.cc/2012/papers/687.pdf.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics, 2016. doi: 10.18653/v1/K16-1002. URL http://www.aclweb.org/anthology/K16-1002.

Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996. URL http://papers.nips.cc/paper/1144-factorial-hidden-markov-models.pdf.

G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995. URL http://www.gatsby.ucl.ac.uk/~dayan/papers/hdfn95.pdf.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. URL http://arxiv.org/abs/1312.6114.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1278–1286, 2014. URL http://jmlr.org/proceedings/papers/v32/rezende14.pdf.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014. URL http://jmlr.org/proceedings/papers/v32/titsias14.pdf.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992. URL https://doi.org/10.1007/BF00992696.