# Unveiling mode-connectivity of the ELBO landscape

**Edith Zhang**
Department of Applied Mathematics
Columbia University
New York, NY 10027
ejz2120@columbia.edu

**David Blei**
Departments of Statistics & Computer Science
Columbia University
New York, NY 10027
david.blei@columbia.edu

## Abstract

We demonstrate and discuss mode-connectivity of the ELBO, the objective function of variational inference (VI). Local optima of the ELBO are found to be connected by essentially flat maximum energy paths (MEPs), suggesting that optima of the ELBO are not discrete modes but lie on a connected subset in parameter space. We focus on Latent Dirichlet Allocation, a model commonly fit with VI. Our findings parallel recent results showing mode-connectivity of neural net loss functions, a property that has helped explain and improve the performance of neural nets. We find MEPs between maxima of the ELBO using the simplified string method (SSM), a gradient-based algorithm that updates images along a path on the ELBO. The mode-connectivity property is explained with a heuristic argument about statistical degeneracy, which is related to over-parametrization in neural networks. This study corroborates and extends the empirical experience that topic modeling has many optima, providing a loss-landscape-based explanation for the "no best answer" phenomenon experienced by practitioners of LDA.

## 1 Introduction

Topic models are hierarchical statistical models used to discover latent structure in a dataset. Latent Dirichlet Allocation (LDA), introduced by [7], is a simple topic model that considers each document as having unique proportions of $K$ topics, where each topic is associated with a distribution over words in a vocabulary. See Appendix A or [7] for details. The statistical inference problem is to optimize the parameters of the model, which is often done using variational inference (VI). In Bayesian inference, one usually seeks the posterior distribution, but in practice it is intractable to compute. VI circumvents this problem by optimizing the parameters of an approximating distribution to be close to the exact posterior. See Appendix B or [15].

There is typically no "best answer" to the task of topic modeling: words take different meanings based on context, and different groupings of words can lead to very different, and yet satisfactory topic assignments [16]. This is the folk intuition of VI performed on LDA, but little is known about the optimization landscape that leads to such results. Though remedies have been proposed, e.g. by allowing users to inject constraints to control the topic outputs [17], there is no formal explanation of the "no best answer" phenomenon.

The objective function of VI is the Evidence Lower Bound (ELBO). It is the sum of an expected log joint and a KL divergence term. To maximize the first term is to find MAP parameters for the model, whereas maximizing the second term is to find an approximate posterior that is close to the prior distribution. VI balances these two goals. However, VI is a bit of a black box: the ELBO is high-dimensional, with a parametrization highly dependent on the model, and thus rather opaque. Gaining a better understanding of the ELBO and its geometry is a key step to our further use of VI in increasingly complex applications.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

We are inspired by the neural net literature, where there has been an increasing fascination in the loss landscapes of neural nets. The ELBO has similarities to neural net loss functions, the goal of both VI and DNNs being to maximize a conditional log likelihood of a hierarchical model with hidden layers. Thus we expect similar properties between the ELBO and neural net losses for analogous models.

One particularly interesting property of neural net loss functions is *mode-connectivity*: recent results indicate that there often exist paths of almost constant loss between local optima. Empirical evidence for mode-connectivity is presented in [11], [12], [14]. Further exploration of neural net loss landscapes and mode-connectivity can be seen in, for example, [12] which relates the nonlinear unit of neural networks to the connectedness of optima, while [8] and [9] suggest that saddle points proliferate in parameter space as dimensions increase. The mode-connectivity property has been exploited to develop new ensembling methods [14]. Furthermore, it is suggested that the MEPs become increasingly flat with wider and deeper neural net architectures [11].

In comparison to neural net loss functions, there has not been much theoretical study of the ELBO. We both provide empirical evidence for mode-connectivity of the ELBO, with a method resembling that of [11]. Our theoretical discussion is closest to [13], who shows that overparametrization and resilience are related to mode-connectivity.

To find MEPs in the ELBO landscape, we use the simplified string method (SSM), introduced by [5]. The SSM is an iterative gradient-based method that sets a straight path between two optimal parameters defined by $n$ beads along it, then alternates 1) a gradient ascent step and 2) a reparametrization step that allow the path to climb to a MEP between local maxima. We apply the string method with a coordinate ascent step which is typical when maximizing the ELBO.

Our theoretical discussion of MEPs draws from analogous discussion on neural net loss functions. It has been shown that overparametrization, i.e. superfluous complexity, leads to near-optimal MEPs in neural networks [13]. We show here that the same holds for the LDA ELBO. In experiments, we show that increased model complexity in LDA also corresponds to increasingly optimal MEPs. However, we argue that there is a more pathological source of mode-connectivity in the ELBO: continuous statistical degeneracy, meaning there is no single optimum but rather a connected set of optimal model parameters. This explains why, as seen in the experimental section, near-optimal MEPs in the ELBO still show up in an underparametrized model.

It is also known that neural net MEPs become increasingly flat with not just wider, but deeper architectures [11]. Analogously, we expect the MEPs to exist (and be even more optimal) in VI for deeper hierarchical models than LDA, which contains only a single hidden layer. Experiments with MEPs in other hierarchical models is left for future work.

Note that the ELBO is to be maximized, whereas neural net losses are to be minimized. In this paper, we refer to the neural net objective functions interchangeably as loss or energy functions. Similarly, MEPs refer to maximum or minimum energy paths in the case of the ELBO or neural net loss function, respectively.

### 1.1 Contribution and Significance

**Contribution** We demonstrate mode-connectivity of the ELBO landscape, paralleling recent results in the neural network literature. The MEPs of the ELBO become increasingly optimal as data size and model complexity increase, indicating that ELBO MEPs, like neural net MEPs, are encouraged by overparametrization. Along the way, this provides an optimization-centric viewpoint of the "no best answer" phenomenon of topic modeling and statistical degeneracy. Practitioners can feel reassured of their results, which are black-box and which may vary based on different initializations.

**Significance** As a black box model, VI and its objective function, the ELBO, are not well-understood. By unveiling mode-connectivity, we seek deeper understanding of the ELBO landscape itself, beyond merely what parameter properties the ELBO encourages, as is done in [22]. Furthermore, by indicating that MEPs abound with increasingly complex models, we conjecture that mode-connectivity holds in deeper probabilistic models. Last, we point out that statistical degeneracy is important: the AI literature often assumes the existence of a single best parameter, but in practice, model misspecification and overparametrization can lead to abundant optima, which may be connected in parameter space. This can lead to unexpected and interesting behavior when it comes time for statistical inference.

## 2 Finding paths between modes

We employ the simplified string method (SSM), introduced by [5], to compute MEPs in the ELBO landscape. The SSM was originally designed to search for chemical transition states, the most probable path of transition between two energetically stable states of a physical system [4],[5]. Applied to the topic modeling inference problem, the potential energy is the ELBO, and optimal transition paths give continuous deformations between two distinct, locally optimal topic configurations. From a statistical degeneracy perspective (see Section 3), these paths occur in connected sets of optimal parameters.

To find a MEP, we first obtain two distinct local maxima of the ELBO using stochastic variational inference (SVI) [23] on the New York Times dataset, which consists of two million text documents. SVI is the stochastic variant of batch variational inference, which alternately updates the global and local parameters in an expectation-maximization fashion (see [15]) using the natural gradient of the ELBO with respect to each of the parameters. SVI gives us optimized global parameters (topics' distributions over words), which are then used to find optimal local parameters (topic proportions for each document) for a smaller held-out dataset. SSM is done on the held-out dataset because the ELBO requires both global and local parameters. See Appendix C for the SSM algorithm.

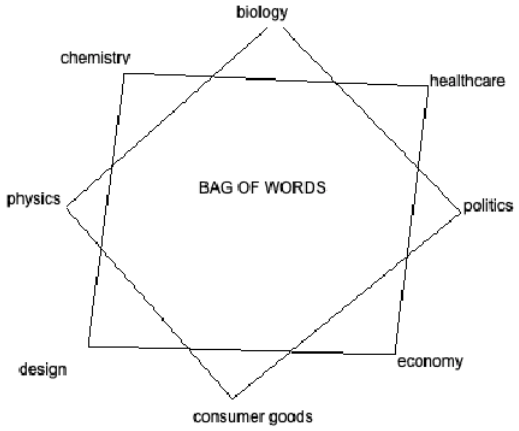
## 3 Explaining MEPs

It is not easy to mathematically prove conditions for the existence of MEPs in neural net losses and the ELBO. Its realization is dependent on the statistical model and the data itself.

However, there exist heuristic arguments explaining the existence of MEPs in neural networks, as discussed by [11], [13]: in an overfitted model, perturbing a single parameter may be "made up for" by changes in the many other parameters, a resilience that is exploited by Dropout [3] and ensembling [14].

The topic modeling analogue would be the following: re-assigning a single word to a different topic corresponds to a slight shift in the topics' identities themselves. As seen in the diagram below, a set of topics can be continously reconfigured to a different set of topics while retaining its descriptive power. The overfitted setting in LDA corresponds to specifying $K > K^*$, or a larger number of topics in the LDA model than actually exist in the corpus. In truth there is no "true number of topics" $K^*$, but our experiments with synthetic data indicate that MEPs are more optimal as the number $K$ of topics in the model grows.

This lends a straightforward analogy to overfitting in neural networks, in which $K > K*$ corresponds to overfitted LDA.




## References

[1] Nguyen, XuanLong. "Posterior contraction of the population polytope in finite admixture models." Bernoulli 21.1 (2015): 618-646.

[2] Tang, Jian, et al. "Understanding the limiting factors of topic modeling via posterior contraction analysis." International Conference on Machine Learning. PMLR, 2014.

[3] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

[4] Weinan, E., Weiqing Ren, and Eric Vanden-Eijnden. "Simplified and improved string method for computing the minimum energy paths in barrier-crossing events." Journal of Chemical Physics 126.16 (2007): 164103.

[5] Weinan, E., Weiqing Ren, and Eric Vanden-Eijnden. "String method for the study of rare events." Physical Review B 66.5 (2002): 052301.

[6] Sheppard, Daniel, Rye Terrell, and Graeme Henkelman. "Optimization methods for finding minimum energy paths." The Journal of chemical physics 128.13 (2008): 134106.

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

[8] Dauphin, Yann, et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization." arXiv preprint arXiv:1406.2572 (2014).

[9] Choromanska, Anna, et al. "The loss surfaces of multilayer networks." Artificial intelligence and statistics. PMLR, 2015.

[10] Bray, Alan J., and David S. Dean. "Statistics of critical points of Gaussian fields on large-dimensional spaces." Physical review letters 98.15 (2007): 150201.

[11] Draxler, Felix, et al. "Essentially no barriers in neural network energy landscape." International conference on machine learning. PMLR, 2018.

[12] Freeman, C. Daniel, and Joan Bruna. "Topology and geometry of half-rectified network optimization." arXiv preprint arXiv:1611.01540 (2016).

[13] Kuditipudi, Rohith, et al. "Explaining landscape connectivity of low-cost solutions for multilayer nets." arXiv preprint arXiv:1906.06247 (2019).

[14] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." arXiv preprint arXiv:1802.10026 (2018).

[15] Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

[16] Boyd-Graber, Jordan L., Yuening Hu, and David Mimno. Applications of topic models. Vol. 11. now Publishers Incorporated, 2017.

[17] Hu, Yuening, et al. "Interactive topic modeling." Machine learning 95.3 (2014): 423-469.

[18] Betancourt, Michael. “Identity Crisis” June 2020, https://betanalpha.github.io/assets/case_studies/identifiability.html.

[19] Wallach, Hanna M., David M. Mimno, and Andrew McCallum. "Rethinking LDA: Why priors matter." Advances in neural information processing systems. 2009

[20] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101.suppl 1 (2004): 5228-5235.

[21] George, Clint P., and Hani Doss. "Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model." J. Mach. Learn. Res. 18 (2017): 162-1.

[22] Hoffman, Matthew D., and Matthew J. Johnson. "Elbo surgery: yet another way to carve up the variational evidence lower bound." Workshop in Advances in Approximate Bayesian Inference, NIPS. Vol. 1. 2016.

[23] Hoffman, Matthew D., et al. "Stochastic variational inference." Journal of Machine Learning Research 14.5 (2013).

# A Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a hierarchical, latent-variable model which aims to discover the hidden semantic structure in a collection of text documents [7]. LDA supposes that there are $K$ topics in a set of documents, and that each document is made up of a different proportion of topics. Each $k^{\text{th}}$ topic is associated with a vector of word proportions, $\beta_k$, sampled from a Dirichlet distribution over the vocabulary, and each $i^{\text{th}}$ document is associated with a vector of topic proportions, sampled from a Dirichlet distribution over the topics.

The LDA generative model is as follows:

1. For each topic $k = 1, \ldots, K$: draw word-proportions $\beta_k \sim Dir_V(\eta)$

2. For each document $i = 1, \ldots, D$: draw topic-proportions $\theta_i \sim Dir_K(\alpha)$
3. For each word $j = 1, \ldots, M_i$ in document $i$:
   (a) Draw assignment $z_{ij} \sim \text{Mult}(\theta_i)$
   (b) Draw word $x_{ij} \sim \text{Mult}(\beta_{z_{ij}})$

The observed data are the words $x_{ij}$ for $i = 1, \ldots, D$. Equivalently, step 3 could be described as a generation of count-data, which is more computationally efficient and which we use in implementation:

3. For document $i$ and vocabulary word $v$: draw word counts $c_i v \sim \text{Mult}(p_{iv})$

where $c_{iv}$ denotes the number of times word $v$ appears in document $i$. Here, $p_{iv}$ is the probability that a given word in document $i$ is word $v$, and is given by $p_{iv} = \sum_k \theta_{ik}\beta_{kv}$. Note that the latents $z$ are marginalized out in this case.

As the name suggests, the inference problem for LDA is to allocate the Dirichlet parameters to the latent variables, which are the document-specific topic proportions and the topic-specific word proportions. Inference is often done with variational inference (VI), outlined in Appendix B.

## B Variational Inference

The central goal of Bayesian inference typically involves maximizing the posterior distribution for a statistical model and dataset. However, computing the posterior is often intractable.

Variational inference (VI) is an inference algorithm that aims to approximate the posterior distribution of a given model and data, by finding the closest distribution to the posterior out of a given family of distributions $\mathcal{Q}$ [15]. The distance from the posterior is measured by the Kullback-Leibler (KL) divergence, or relative entropy, defined between probability distributions $\mu$ and $\nu$ to be

$$KL(\mu||\nu) \int \log\left(\frac{\mu(x)}{\nu(x)}\right)\mu(dx).$$

Thus the objective of VI is to find

$$q^*(\beta, \theta, z) = \arg\min_{q \in \mathcal{Q}} KL(q(\beta, \theta, z)||p(\beta, \theta, z|x)). \tag{1}$$

The objective function $KL(\cdot||p(\beta, \theta, z|x)$ can be rearranged using Bayes' rule, and removing constant terms [15] to give the equivalent objective which is called the ELBO:

$$\mathcal{L}(\lambda, \gamma, \phi) = \mathbb{E}_q[\log p(\beta, \theta, z, x)] - \mathbb{E}_q[\log q(\beta, \theta, z)]. \tag{2}$$

The first term is the expectation of the LDA log joint with respect to the variational distribution $q$, and the second term is the negative entropy of $q$. Thus VI encourages distributions $q$ that yield a high expectation of the joint and have high entropy (spread).

In practice, $\mathcal{Q}$ usually chosen to be the mean-field family, which assumes variables are independent. Thus distributions $q \in \mathcal{Q}$ take the product form

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k; \lambda_k) \prod_{i=1}^{D} q(\theta_i; \gamma_i) \prod_{i=1}^{N} \prod_{j=1}^{M_i} q(z_{ij}; \phi_{ij}).$$

In other words, the mean-field family consists of distributions $q$ in which all the topics $\beta$, documents' topic proportions $\theta$, and word topics $z$ are independent of each other. The parameters $\lambda, \gamma$, and $\phi$ are called the *variational parameters*, which are to be optimized over. They take the forms $\lambda \in^{K \times V}, \gamma \in^{D \times K}$, and $\phi \in^{D \times V \times K}$, with interpretations as follows.

$\lambda_k \in^V$ is the parameter governing topic $k$, and is a vector of frequencies of words. It is a $V$-vector, where the vocabulary is of size $V$. The mean-field distribution posits that the topic $\beta_k$ is a Dirichlet($\lambda_k$) random variable. This is of the same form as the LDA model, except with the assumption of independence. Specifically,

$$q(\beta_k; \lambda_k) = \text{Dir}(\lambda_k).$$

Similarly, the parameter governing document $i$ is $\gamma_i$, which are proportions of the $K$ topics in document $i$:

$$q(\theta_d; \gamma_d) = \text{Dir}(\gamma_d).$$

The observed variables are governed by the parameter $\phi$, so that $\phi_{ijk}$ is the probability that word $j$ in document $i$ is assigned to topic $k$:

$$q(z_{ij}; \phi_{ij}) = \text{Cat}(\phi_{ij})$$

or equivalently in the case of count data,

$$q(c_{iv}; \phi_{iv}) = \text{Mult}(\phi_{iv}, M_i).$$

Though the independence assumption seems strong, it is flexible., and a reasonable assumption in topic modeling.

The variational parameters are optimized according to a coordinate ascent algorithm [15] or a stochastic variational inference (SVI) which uses stochastic natural gradient ascent [23]. In our experiments, we opt for SVI which is scalable to massive dataasets.

## C Description of the simplified string method

Two runs of SVI from random initializations give two distinct optimal global parameters $\lambda_1$ and $\lambda_2$. Batch VI is run for the same held-out dataset to obtain optimal local parameters with respect to each of the optimal global parameters, to obtain two optima $m_1 = (\lambda_1, \gamma_1, \phi_1)$ and $m_2 = (\lambda_2, \gamma_2, \phi_2)$ in parameter space.

$N$ equally-spaced points ("beads") between $m_1$ and $m_2$ are interpolated. Each natural gradient step on a bead $(\lambda, \gamma, \varphi)$ is equivalent to a batch VI update on all three parameters.

The SSM algorithm:

1. Initiate $N = 15$ beads equally spaced along the line segment between two optima.
2. Gradient step: each bead takes a natural gradient ascent step, which has a closed form for exponential family models.
3. Reparameterization step: to prevent beads from running off to maxima, slide them along the string so they are equally-spaced along their piecewise linear path.
4. Repeat steps 3 and 4 until convergence.