Analysis of ELBO Loss Landscape: No Barrier In Between Local Maxima

Ang Li[*]
Columbia University
New York
al4263@columbia.edu

Xinyu Wei[*]
Columbia University
New York
xw2762@columbia.edu

Abstract

We present a comprehensive empirical study of the Evidence Lower Bound (ELBO) landscape in variational inference, addressing fundamental questions about the geometric properties of its local maxima and their connectivity. While prior work has suggested the existence of Minimum Energy Paths (MEPs) connecting local optima in the ELBO parameter space, systematic empirical validation remains incomplete. We adapt the string method to construct MEPs and introduce a novel visualization approach using random vector projection to analyze the ELBO landscape. Through extensive experiments on Latent Dirichlet Allocation with the AP dataset, we (1) provide the first empirical evidence for the existence of continuous low-energy paths between local maxima, (2) demonstrate that these paths preserve semantic interpretability of topics, and (3) quantify the generalization performance of solutions along MEPs. Our findings reveal previously unknown geometric properties of the ELBO landscape and provide practical insights for initialization strategies in variational inference.

1 Introduction and Motivations

In variational inference, one maximizes the ELBO using CAVI. Due to the non-convex nature of the ELBO optimization objective, CAVI converges to a local maximum sensitive to initialization. In this project, we investigate the landscape of the ELBO by empirically analyzing the relationship between local maxima obtained in a hierarchical graphical model. This will help VI practitioners to interpret the different local maxima that CAVI reaches, e.g. many local maxima are symmetrical and induced due to label switching of cluster assignment[1]. We are intrigued by the questions such as whether a linear interpolation of local maxima gives other local maxima and whether there are other geometric properties between these local maxima apart from their symmetry. It has been previously suggested that in the parameter space of ELBO, there is a connected region of low objective value, also known as Minimum Energy Paths (MEPs), between two local maxima[1]. Nevertheless, the experimental results of this phenomenon of mode connectivity in ELBO are not complete. In this project, we:

1. adapt and implement the string method proposed by [2] to find the Maximum Energy Paths (MEPs) in the ELBO loss landscape, thereby validating the hypothesis of their existence;
2. provide a visualization of the ELBO loss landscape using random vector projection;
3. empirically investigate the interpretations of the MEPs beyond heuristic explanations; and
4. evaluate whether the points along the MEPs perform well on out-of-sample data.

We use the LDA model with the AP dataset for this project due to computational limits (we will further discuss the choice of model and dataset in the criticism section). In the following section,

[*]Equal contribution.

Algorithm 1 NEB Maximum Energy Path Construction for ELBO

Input: initial path $p^{(0)}$ with N+2 pivot, γ (learning rate)
$p_0^{(0)} \leftarrow \theta_1$
$p_{N+1}^{(0)} \leftarrow \theta_2$
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, N$ **do**
    Compute perpendicular force $F_i^L|_\perp$
    Compute horizontal force $F_i^S|_\parallel$
    $F_i \leftarrow F_i^L|_\perp + F_i^S|_\parallel$
    $p_i^{(t)} = p_i^{(t-1)} + \gamma F_i$
  end for
end for
return final path $p^{(T)}$

we will present the algorithm adapted from the string method to find the MEPs in the ELBO loss landscape and the method for visualizing the landscape. Then, we will demonstrate the result of MEP construction with a visualization of the ELBO loss landscape. In particular, we show the results of the MEPs and loss landscape of different initializations with different hyperparameters. We will discuss the results in accordance with the topic discovered and the most probable words to illustrate the concept of label-switching proposed by [1]. Our code can be found on the GitHub link Here.

2 The Model, ELBO and Maximum Energy Path (MEP)

The maximum energy path describes the path with the highest objective (ELBO) value between two local extremal points of the optimization objective function (ELBO). We use the following method to obtain such a path.[1][2]

First, we find two sets of variational parameters denoted as $\theta_1 = (\lambda_1, \gamma_1, \phi_1)$ and $\theta_2 = (\lambda_2, \gamma_2, \phi_2)$ that maximize the ELBO with different initializations using CAVI. To find the MEP between them, we assume there exists a continuous path $p*$ with high ELBO values between the two sets of variational parameters θ_1, θ_2. In low-dimensional space, such a path is easy to construct using dynamical programming. For high-dimensional space, such an approach is infeasible. Instead, we use a method called Nudged Elastic Band (NEB) to approximate the path. The state of art method was inspired by a method for connecting extrema in statistical mechanics. We adapted the method for finding the MEPs for ELBO. Suppose the path between θ_1 and θ_2 can be approximated using $N + 2$ pivots $p_0, \ldots, p_{N+1}$, where $p_0 = \theta_1$ and $p_{N+1} = \theta_2$. The goal is to find the path through these pivots that minimizes the energy function $E(p)$ (which in our case is the negative ELBO objective plus the spring energy):

$$E(p) = \sum_{i=1}^{N} L(p_i) + \sum_{i=0}^{N} \frac{1}{2} k \|p_{i+1} - p_i\|^2$$

where $L(p_i) = -\text{ELBO}(p_i)$ and k is the spring constant that prevents the adjacent pivots from stretching too far apart. This minimization problem will produce a path that has low energy and is relatively smooth.

The NEB method uses forces, which are the gradient of the negative ELBO objective with respect to each pivot parameter, to update the parameters of the pivots, thus minimizing the negative ELBO (maximizing ELBO) on the path. The force at each pivot can be divided into the loss force and the spring force: $F_i = -\nabla_{p_i} E(p) = F_i^L + F_i^S$. In our case, F_i^S is simply $k \cdot (p_i - p_{i+1})$, and the F_i^L is the gradient of negative ELBO with respect to p_i: $\nabla_{ELBO}(p_i)$. To intuitively understand these two forces, the spring force moves the pivots closer to each other whereas the loss force moves the pivots in direction of maximum increase of their ELBO values. These together minimize the total energy $E(p)$.
To minimize the energy, the NEB "nudge" the force so that the loss force acts perpendicularly to the

path and the spring force only acts in the parallel direction of the path. And the update rules are:

$$F_i^L|_\perp == -(\nabla_{ELBO}(p_i) - (\nabla_{ELBO}(p_i) \cdot \hat{\tau}_i))$$
$$F_i^S|_\parallel == (F_i^S \cdot \hat{\tau}_i)\hat{\tau}_i$$

where $\hat{\tau}_i$ is defined as:

$$\hat{\tau}_i = \mathcal{N} \begin{cases} p_{i+1} - p_i, & \text{if } L(p_{i+1}) > L(P_i - 1) \\ p_i - p_{i-1}, & \text{else} \end{cases}$$

where $\mathcal{N}$ normalizes the resulting parameter vector. The forces act as "gradients", and the update rule for the NEB method is simple gradient descent with learning rate $\lambda : p_i(t) = p_i(t-1) + \lambda F_i$. Algorithm 1 describes the full algorithm.

3 ELBO Loss Landscape Visualization

To visualize the ELBO loss landscape, we adapt the visualization method for neural networks proposed by[3] . To visualize the high-dimensional variational parameter space, we use a random projection method to project points in high-dimensional space into 2-dimensional space and graph the ELBO values of the different variational parameters in the low-dimensional space. We find two random orthonormal vectors to form the basis of the 2-d space and then project the parameters onto this 2-d space. To illustrate this, we can see Figure 1 in which we have two random orthogonal directions u and v and a center point w_1. We obtain the 2-dimensional embedding (x, y) of a high-dimensional variational parameter θ by first centering the parameters at w_1 and then projecting the parameters to u and v respectively. This means that

$$x_{proj} = (\theta - w_1) \cdot u \tag{1}$$
$$y_{proj} = (\theta - w_1) \cdot v \tag{2}$$

We then plot the sets of 2-d embeddings of pivots on the MEP as well as the two local maximums attained θ_1, θ_2. We also plot a loss surface grid on which we evaluate the ELBO corresponding to each point in the grid and plot the pivots on the grid. The grid is constructed by first choosing the center w_1 as the midpoint between θ_1 and θ_2. Then, we construct a matrix with values at (x, y) being the ELBO value evaluated at $w_1 + xu + yv$.

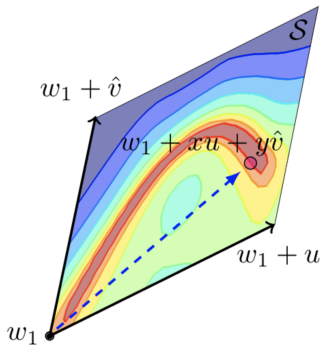


Figure 1: Loss landscape visualization method

4 Experimental Results and Discussion

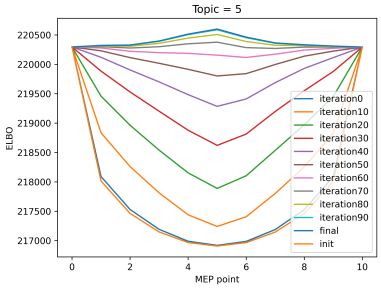
In this section, we present the experimental results where we finetuned the hyperparameters and discuss the effects of hyperparameters and other relevant findings.

4.1 Finding Local Maxima.

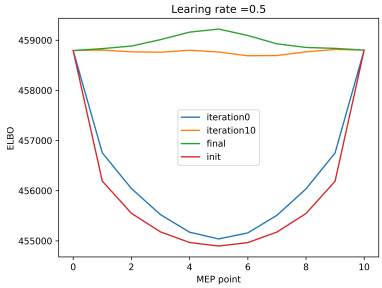
We ran two independent CAVI with independent starting points to find two sets of local maxima to be the two end points of the MEP. As shown in the appendix, we fixed the number of topics and ran CAVI until convergence.

4.2 Path Construction and Hyperparameters.

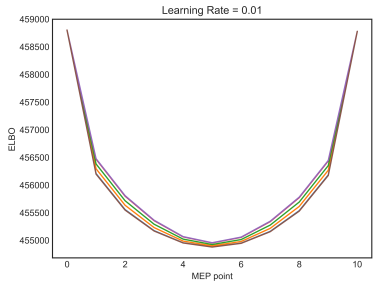
For the path construction, we tested on the AP dataset with three different learning rates: $0.5, 0.1, 0.01$ with a spring constant of 0.01. For learning rates of 0.1 and 0.5, the NEB method encounters some numerical issues at around 30 epochs and 10 epochs respectively, which is when we stopped the training (seen in Figure 2 (a) and (b)). Nevertheless, due to the large learning rate, the model constructed paths with clearly high ELBO values than the linear interpolation. For a learning rate of 0.01, the model makes little progress on finding a path with low energy, even with 100 iterations (seen in Figure 2 (d)). The best learning rate we figured out is $lr = 0.1$ shown in Figure 2 (a) which shows a relatively flat path among the other plots. From the graphs, we can see that the pivots on the path have higher ELBO values after each iteration and the ELBO value approaches those of the two local maxima θ_1, θ_2.



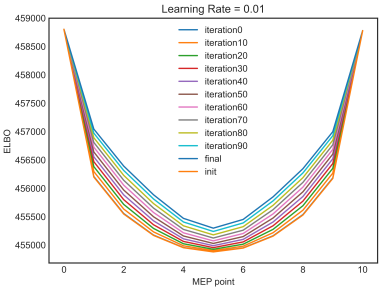
(a) lr = 0.1 with 100 iterations



(b) lr = 0.5 with 17 iterations



(c) lr = 0.01 with 30 iterations

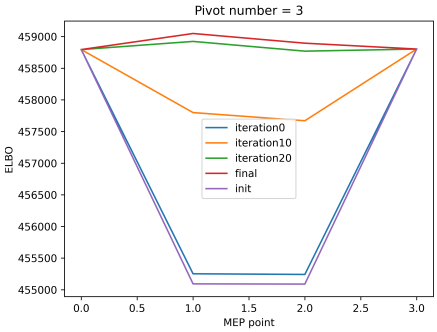


(d) lr = 0.01 with 100 iterations

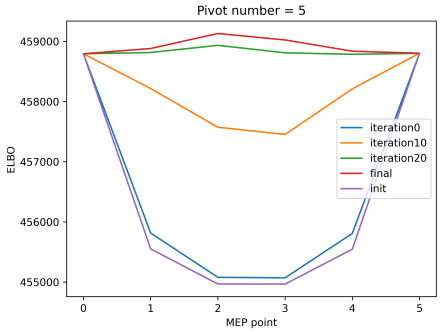
Figure 2: MEP path construction and the ELBO values using different learning rates

From the above plot, we can see that the constructed final path is a path with much higher ELBO values than the linear interpolation of points in the parameter space. The results match our intuitions and serve as a numerical validation that there is no barrier between two local maxima in the ELBO landscape.

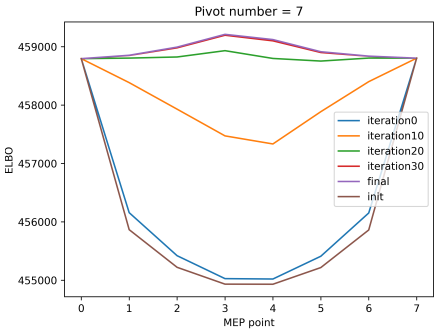
For the number of pivots, we tested on $N = 3, 5, 10, 20$ (seen in Figure 3). For all N shown in Figure 3 (a), we obtained an almost linear path and found points that have a higher ELBO value than the initial local maxima. To systematically study the ELBO path generation, we chose $N = 10$ for the



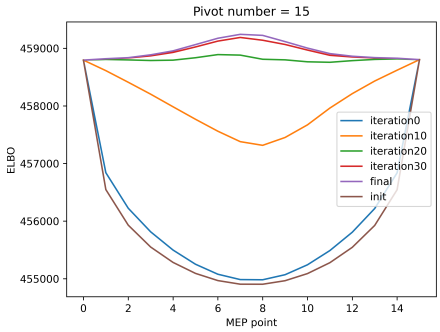
(a) Results for pivot count N = 3



(b) Results for pivot count N = 5



(c) Results for pivot count N = 7



(d) Results for pivot count N = 15

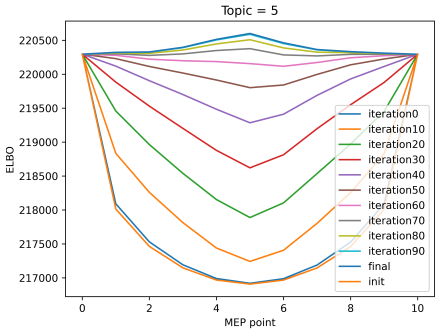
Figure 3: MEP path construction and the ELBO values using different number of pivots

other following experiments. For the spring constant, we tested with K = 0.1 0.005, and 0.001. What is surprising is that during the MEP construction, the algorithm even found parameters that have a higher ELBO value than the two maxima found by CAVI as shown in Figure 5 (a). This suggests that the MEP algorithm may be a way to find new local maxima.

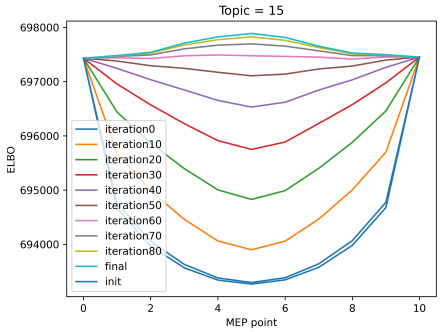
We also tested the different number of topics C assumed for the LDA model. The results suggest that the number of topics does not influence the construction of MEP but the corresponding ELBO values are different as shown in Figure 4. Nevertheless, the highest log posterior predictive occurs at $K = 20$ as shown in the appendix. This can be interpreted as evaluating the goodness of fit of the model and that when $K = 20$, the model is more able to make predictions of words for unseen parts of the out-of-sample data. This is reasonable since when the number of topics is too small, it may not be able to capture the distinct themes across documents and may be too generic. As the number of topics increases, more distinctive themes may arise and the resulting model becomes more capable of capturing the common themes.

4.3 Quality of parameters

To evaluate the quality of the sets of parameters found along the MEPs, we tested the predictive likelihood score on both the training data set and the test data set (seen in Figure 6). To our surprise, while the final MEP pivots have increased performance on the training dataset, they have a smaller log predictive likelihood. This means that the trained pivots have worse generalization abilities than untrained ones. This is surprising since it suggests that the linear interpolation between the two local maxima has even better generalization abilities on the held-out dataset. We have not yet found explanation for this.

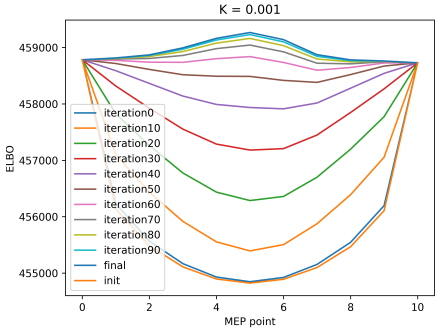


(a) topic = 5

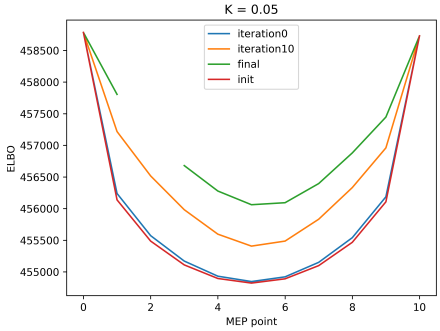


(b) topic = 15

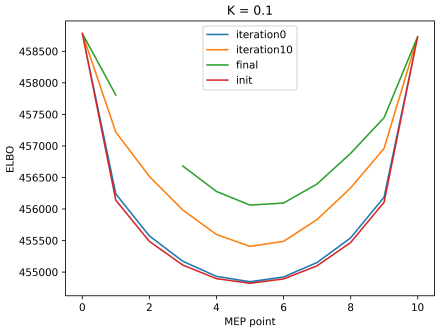
Figure 4: MEP construction and effects of different number of topics



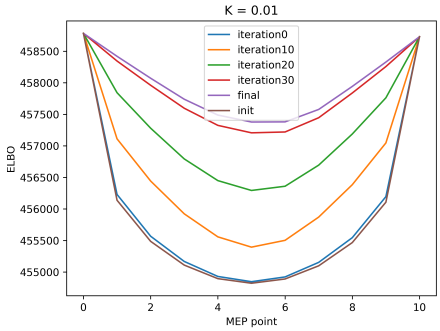
(a) k=0.001



(b) k = 0.005

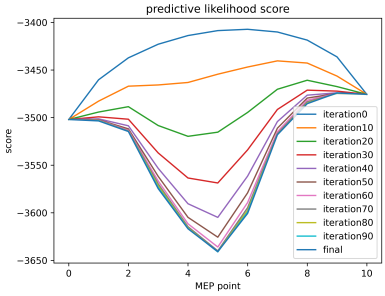


(c) k = 0.1

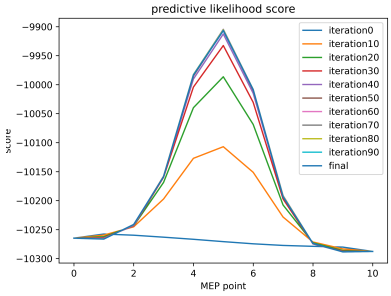


(d) k = 0.01

Figure 5: MEP construction and different string constant k



(a) predictive score of θ on the path for testing data

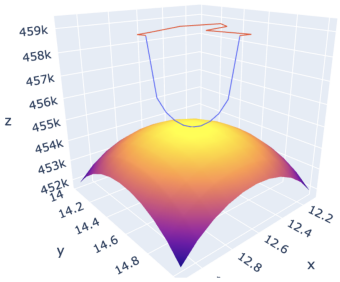


(b) predictive score of θ on the path for training data

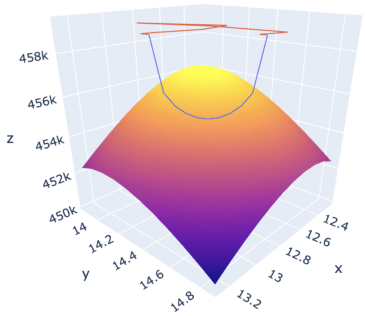
Figure 6: Predictive score of parameters on the MEP

4.4 Visualization

To visualize the ELBO landscape with the method above, we used two different approaches to find the set of orthogonal random vectors as a basis for projection: the first one is to draw two vectors $\in R^d$ from random multivariate Gaussian distributions. The other approach is to generate two orthogonal vectors with a high-dimensional Bernoulli. The results are as follows.



(a) Gaussian random vector projection landscape



(b) Orthogonal vector projection landscape

Figure 7: Visualization of the ELBO loss landscape where the red curve is the MEP found using the path construction and the blue curve is the interpolation between θ_1, θ_2

Here, the red line is the constructed path between the two local maxima whereas the blue line is a linear interpolation between two local maxima. It is clear that the red line is above the blue line and the landscape, which means the path we generated is a "max ELBO" path that connects the two local maxima. However, the loss landscape (the loss on the grid) did not quite match with the path and the linear interpolation, which may be that multiple high-dimensional vectors are projected onto the same point in the 2-dimensional space. Moreover, the non-convexity of the landscape was not captured by the grid which may be due to that when constructing the grid, we only sampled a local region around the center of the grid which may correspond to a basin in the actual ELBO landscape. To better capture the landscape of ELBO, we need to reconsider the scale at which we construct the grid.

4.5 Interpretation of topics and words

To interpret the sets of parameters along the MEPs, it is interesting to look at the topics and words discovered along the path. We can use the fitted variational parameters to approximate the posterior topics $\mathbb{E}[\beta_k|x]$. Here, we choose two pivots along the path and find the most probable words of each topic, and compare them with the words of topics from two initial local maxima. For columns 1 and 2 of θ_1, the topics can be summarized as political economics and military actions. If we look at the

first column of pivot 1, it is mostly about bills and military actions, which is the combination of the first two columns of θ_1. This corresponds to Blei's (2018) argument about label switching between sets of parameters found by the ELBO optimization. Another example is that the fourth column in pivot 2 is about federal actions on federal bills, which is split into education and bills in column 9, column 3, and column 8 in θ_2. This can be summarized as non-identifiability, which means that the topics are fundamentally non-separable and thus exists many ways to summarize them.

six	country	go	world	believe	dukakis	states	united	saturday	state
time	military	two	first	officials	year	soviet	last	leaders	bank
statement	duracell	children	american	agreement	company	york	embassy	like	barry
end	died	fire	monday	friday	price	year	people	three	tuesday
take	meeting	congress	defense	union	asked	northwest	head	earlier	today
area	work	man	police	billion	two	week	officers	new	california
campaign	administration	made	thats	bush	told	group	people	official	i
economy	news	prices	rose	federal	oil	report	government	percent	new
noriega	mrs	security	wednesday	high	government	city	states	central	president
president	economic	national	new	business	year	rate	years	percent	i

(a) Most probable words from topics according to θ_1

senate	go	saturday	officials	six	believe	state	states	united	dukakis
military	last	bank	years	two	year	first	officials	soviet	leaders
agreement	statement	duracell	barry	embassy	like	new	york	company	children
billion	four	fire	office	friday	price	year	people	monday	three
take	day	northwest	asked	head	defense	earlier	today	congress	union
california	billion	car	area	scientists	officers	two	new	man	police
thats	campaign	told	president	made	group	i	people	official	bush
prices	government	federal	million	new	economy	rose	oil	percent	report
noriega	high	president	rating	new	john	wednesday	government	security	central
dont	see	month	national	economic	business	rate	years	percent	i

(b) Most probable words from topics according to pivot1

group	made	today	top	saudi	defense	union	company	officials	state
new	officials	year	school	last	years	first	soviet	police	two
united	police	west	told	mexico	meeting	monday	people	central	southern
million	people	economy	get	prices	billion	months	i	year	last
national	record	think	federal	increased	business	month	percent	rate	california
global	world	man	service	county	war	waste	fire	make	leaders
news	people	party	states	saying	official	campaign	bush	president	state
forces	people	report	money	new	government	million	program	city	oil
rating	new	president	head	talks	bank	plan	week	northwest	back
good	believe	times	dont	roberts	two	american	economic	dukakis	i

(c) Most probable words from topics according to pivot2

businesses	saudi	today	defense	made	top	group	officials	company	union
died	years	last	police	new	two	first	office	school	soviet
west	mrs	jackson	police	meeting	monday	southern	people	told	united
million	months	economy	going	get	last	year	billion	prices	i
record	increased	california	federal	month	national	think	percent	rate	business
animals	county	service	world	friday	price	leaders	fire	make	war
officers	countries	news	saying	campaign	states	washington	official	president	state
money	military	million	oil	wednesday	program	soviet	people	city	government
time	northwest	week	plan	back	head	man	talks	bank	new
noriega	asked	economic	dukakis	roberts	two	good	american	bush	i

(d) Most probable words from topics according to θ_2

Figure 8: Most probable words from each topic, with K = 10

5 Criticism

There are several drawbacks and potential improvements to our research and experiments: First of all, it is hard to assess the convergence of the NEB. The NEB methods heavily rely on hyperparameter tuning and thus make it difficult to have a convergence criterion. To solve this, we could implement Auto-NEB, which discards the hyperparameter k, and redistributes the pivot every iteration. The method can also automatically add more pivots to deal with the more complex landscape as needed.

Second, when plotting the ELBO landscape with the visualization method, we found that the ELBO of points on the landscape did not quite match the path we constructed. Our assumption is that when embedding a high-dimensional vector into a 2-dimensional vector, there is a great chance that many different high-dimensional vectors are projected onto the same 2-dimensional vector, which causes the mismatch issue. Moreover, as suggested above, the scale at which we graph the grid may be too local which resembles a concave space and does not capture the non-convexity of the ELBO landscape. Thus, we may need to reconsider the scale that we sample points around the center point.

6 Future Works

Based on the explorations and experiments, the next step towards exploring the ELBO landscape is to try different visualization techniques and path construction techniques. A different path construction procedure was proposed by [3] based on Bezier curve parametrization of the path. One can also explore the symmetry structure of the landscape and try to experimentally see whether there exist symmetrical landscapes which may correspond to label switching. Also, it was suggested by [4] that a wider and deeper neural network may produce a flatter region than MEPs and thus we hypothesize that a similar phenomenon may be true for the ELBO loss landscape corresponding to a deeper hierarchical model with more parameters. This can be investigated in future works. Moreover, theoretical analysis of the ELBO loss landscape is beneficial for establishing and formal framework to characterize the ELBO loss and theoretically explain the meaning of different sets of variational parameters.

References

[1] Edith Zhang and David Blei. Unveiling mode-connectivity of the elbo landscape.

[2] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

[3] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[4] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

Appendix

Github link

https://github.com/LeonLixyz/6701_Project

Dataset

AP dataset with each article as a sparse vector of length equal to the size of the vocabulary. The LDA graphical model is as follows.

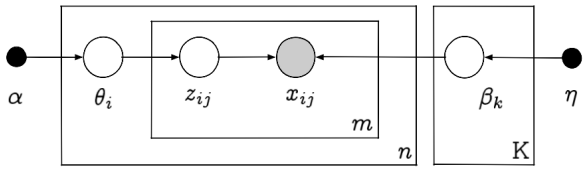


Figure 9: Graphical Model of LDA

Further Details about the Model and Experiments

Data Generation Process and the Posterior. The data is assumed to be generated from a mixed membership model (LDA). We first fix the number of topics and for each topic K, for each k from 1 to K, we draw β_k from a categorical distribution on the vocabulary. Each β_k is a vector representing the probability of each word belong to each topic. These parameters are shared across all documents and using this, we would be able to discover common structure of words which are grouped into different topics. Then, for each document, we draw a topic proportion θ_i from a categorical distribution over all topics. Each vector represents the amount each topic is expressed in the current document. Then, for each word (j) in the current document (i), we draw topic assignment z_{ij} from the categorical distribution which was drawn in the previous step. Lastly, we draw the specific word x_{ij} from a categorical distribution from the β vector of the assigned topic. With this generative model, we compute the log posterior which is

$$\log p(\beta, \theta, z|x) = \sum_{k=1}^{K} \log p(\beta_k) + \sum_{i=1}^{n} \log p(\theta_i) + \tag{3}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (\log p(z_{ij}|\theta_i) + \log p(x_{ij}|z_{ij}, \beta)) - \log p(x) \tag{4}$$

Since this posterior is intractable to compute, we use variational inference to approximate the posterior distribution.

Variational Inference. To approximate the posterior $p(\beta, \theta, z|x)$, we use the variational family $q(\beta, \theta, z; v)$ which is parametrized by the variational parameters v. For this homework, I used the mean-field family which is consistent with the lecture. The goal is to find the optimal v^* such that the difference between the approximate distribution and the actual posterior distribution is minimized. The objective function for minimization is the KL divergence which is

$$KL(q(\beta, \theta, z; v)||p(\beta, \theta, z|x)) \tag{5}$$

Expanding the KL divergence and upon further derivations, we can derive the ELBO and it was shown in the lecture that maximizing the ELBO is equivalent to minimizing the KL divergence, i.e.

$$ELBO = \mathbb{E}_q[\log p(\beta, \theta, z, x)] - \mathbb{E}_q[\log q(\beta, \theta, z; v)] \tag{6}$$

To maximize the ELBO, we use the coordinate ascent variational inference (CAVI) which iteratively updates each variational parameter until convergence of the ELBO objective.

Predictive Score. We used the generalization idea about the completion which is that given some observed words from the held-out dataset and the in-sample data, can we predict the unobserved words? The log predictive probability of the unobserved words indicates the generalization ability of the model. The partially-observed predictive score is calculated as

$$\Lambda^{partial}(K) = \sum_{i=1}^{n_{out}} \log p(x_{out,i}^{new} | x_{out,i}^{obs}, x_{in}; K) \tag{7}$$

We used this criterion to evaluate the model and hyperparameters.

Convergence of ELBO Below are some of the ELBO convergence plots generated when finding the two maxima points.

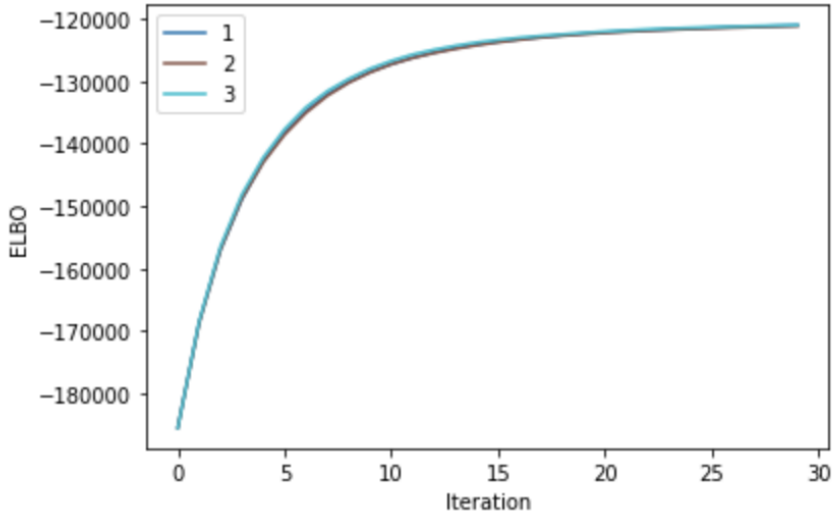


Figure 10: Convergence of ELBO objective with 3 restarts

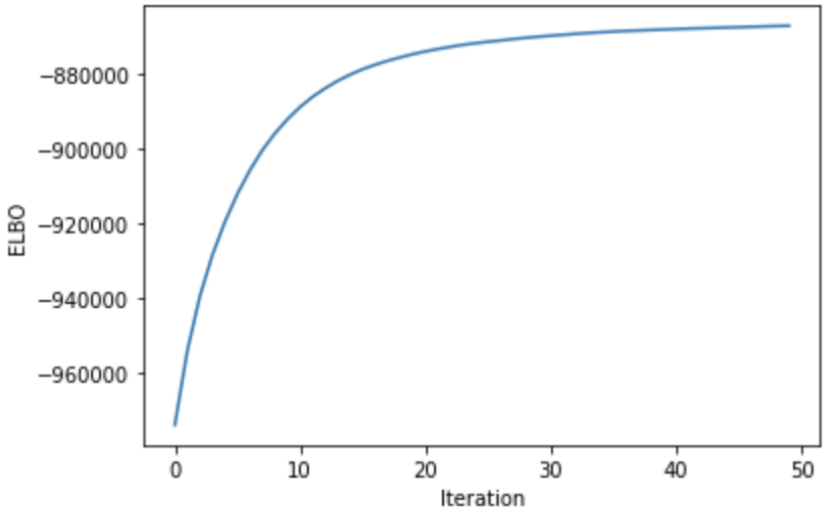


Figure 11: ELBO convergence with K = 5

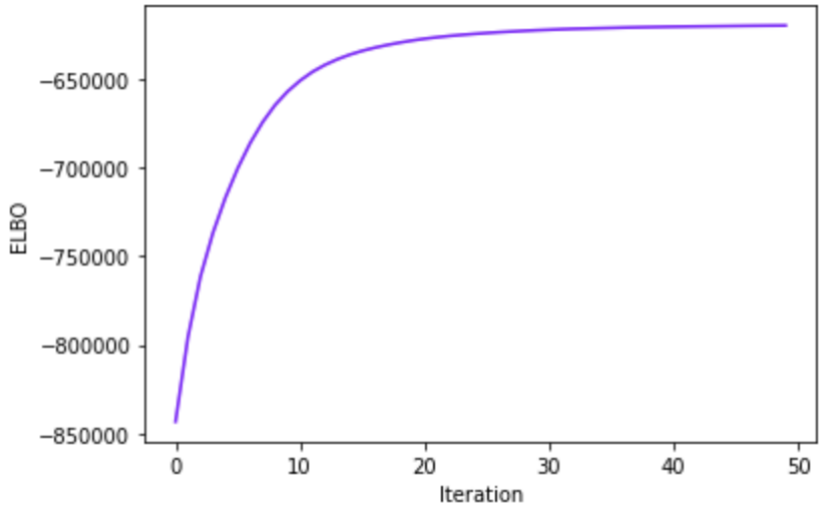


Figure 12: ELBO convergence with K = 10

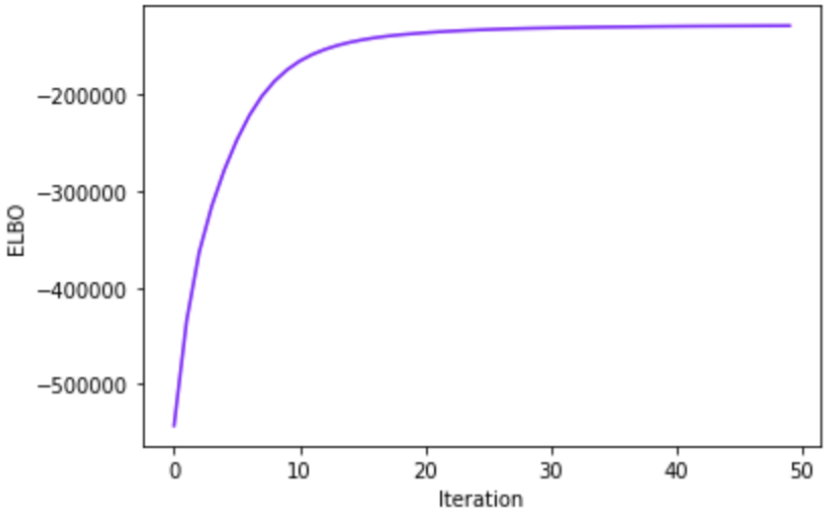


Figure 13: ELBO convergence with K = 15

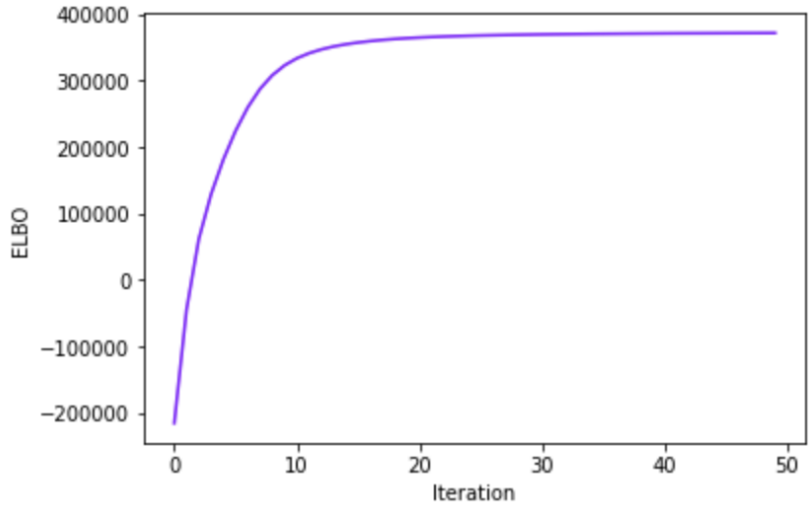


Figure 14: ELBO convergence with K = 20

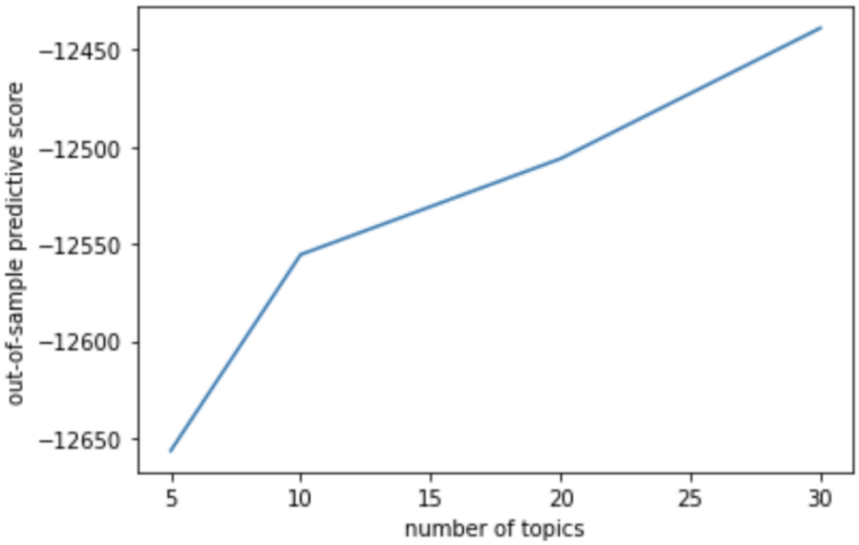


Figure 15: Out-of-sample predictive score and number of topics