Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MASTERTHESIS

Variational Inference for Scalable Probabilistic Topic Modelling

vorgelegt von

Kai Brusch

MIN-Fakultät

Fachbereich Informatik

Language Technology Group

Studiengang: Intelligent Adaptive Systems

Matrikelnummer: 6886748

Erstgutachter: Prof. Dr. Chris Biemann

Zweitgutachter: Steffen Remus

Abstract

Topics models allow to summarize and categorize vasts amount of information. Latent Dirichlet Allocation (LDA) is a well-established topic model that uses a probabilistic approach to estimating distributions over words, documents, and topics. LDA has been shown to have two problems (1) scalable estimation and (2) poor semantic coherence of topics. Probabilistic models are notoriously complex to estimate, often relying on sampling based methods. Variational Inference presents an algorithm that allows circumventing this problem by posing the estimation as an optimization problem. Attention-Based Aspect Extraction has been introduced as an improvement over LDA by enforcing topic coherence with an attention mechanism. This thesis investigates the link between estimation algorithm and resulting semantic coherence of LDA and ABAE.

Contents

Chapter 1

Introduction

1.1 Background

> We are drowning in information but starved for knowledge
> — John Naisbitt

The advent of the Information Age has started to transform every aspect of our lives into a sequence of ones and zeroes. The amount of digital information has been growing exponentially for the past year, while at the turn of the century Gigabytes seemed large we are now thinking in Exabytes. The internet and modern sciences are examples of areas that produce previously unimaginable amounts of data. A current estimate gauges that video streaming platform Youtube will produce 1–2 exabytes of video by 2025 (1 exabyte is 10^{18} bytes), genome researchers have surfaced an estimate that by 2025, between 100 million and 2 billion human genomes could have been sequenced. The data-storage demands for this alone could run to as much as 2–40 exabytes [Check Hayden, 2015]. This glut of information has the potential to fundamentally improve medicine, government, and all scientific research. The underlying assumption is that we can understand more about the world around is by examining data and seek to explain it with models. Though I fundamentally I agree with this assumption I see a gap between the excess of data and progress. Data alone will not spark innovation or progress, it will be models transforming terabytes into new knowledge. Without a model, data is just a binary representation of an observation. There exists a wide range of models for extracting these patterns. The true marvel of the Information Age is the ability to discover hidden patterns in these vast space of data. Thus the problem lies in two aspects: how to capture the underlying systems and how do scale this up to terabytes of data.

The difference between data and information is a model. A model is a certain interpretation of the data that seeks to explain the generating process. The method of finding hidden patterns in data is commonly referred to as Machine Learning. Machine Learning is a statistical framework to develop, explore and evaluate models which explain data. A model applied to data will produce one possible explanation, however, there are many competing models to explain a single phenomenon. Machine Learning provides techniques to produce, evaluate or find competing explanations. Traditionally Machine Learning Models can be broadly be

categorized into supervised and unsupervised. Supervised models assume that the data set is annotated with a true observed label, formally this means that the data $\mathcal{D}$ consists out of inputs x and output labels y. At the heart of supervised learning stands the data representation $\mathcal{D}_{supervised} = \{(y_i, x_i)\}_{n=1}^{N}$, where N is the number of training samples. Supervised learning has been extremely successful, the most known member is the linear regression which is the backbone of science. On the other side of the spectrum, are unsupervised models. Unsupervised models include clustering and topic modeling. As these model's name suggests the aim is to find topics or clusters which group data together. Finding these labels stands at the heart of unsupervised learning. The corresponding training set for unsupervised can be formalized as $\mathcal{D}_{unsupervised} = \{(x_i)\}_{n=1}^{N}$. The central difference is that unsupervised learning does not have the associated label y for each data point x. Is the nature of unsupervised methods to have less information to leverage. Less information allows for a large space of solutions in which the labels lie. This loose nature of this unsupervised models lends itself to more to an exploratory approach and also referred to as knowledge discovery [Murphy, 2013]. In the absence of labels for $\mathcal{D}$ the problem of unsupervised modeling is expressed as a density estimation problem. Clustering as a density estimation problem is formalized as $p(x_i|\theta)$, unlike the supervised case, which is formalized as $p(y_i|x_i, \theta)$. These formalizations highlight supervised learning as conditional density estimation and unsupervised as unconditional density estimation. With this formulation of unsupervised learning, the main question of the field can be formulated as: what set of parameters are optimal, what is optimal and how to find the parameters? The answers to that question can be categorized again into two approaches: Frequentist and Bayesian.

At the theoretical heart of machine learning lies the understanding of unknown quantities. The fundamental principles of understanding these quantities can be done in two frameworks: Frequentist and Bayesian (Probabilistic). The frequentist approach uses point estimates of the unknown quantities. Frequentist only ever conditions on the actually observed data, in this approach there is no notion of repeated trials [Murphy, 2013]. For the clustering example from the previous section, a frequentist would argue in the following fashion. There exists a likelihood function $p(X|\theta)$ for which we are trying to infer the θ value that maximises the likelihood. This approach is called Maximum Likelihood Estimation (MLE) is the standard estimation technique for Frequentist statistics.

$$\begin{aligned} \theta_{MLE} &= \arg\max_{\theta} p(X|\theta) \\ &= \arg\max_{\theta} \prod_{i} p(x_i|\theta) \end{aligned} \tag{1.1}$$

Though it's much more practical to find the θ value to optimize the log-likelihood we can think of the described approach as 'classical frequentist statistics'. The estimation technique for MLE is usually done with a gradient-based approach. These approaches optimize the gradient of the loss function. Though very common the frequentist approach has been described to be plagued by so-called: pathologies. Pathologies are deficiencies in interpretability of p-values and confidence intervals and violation of the likelihood principles [Murphy, 2013]. Each pathology is a consequence of using point-estimates from the MLE approach outlined. A different approach to the statistical foundations of machine learning is the probabilistic or

Bayesian approach. The Bayesian approach is interested in the full distribution of unknown quantities rather than point estimates and uses the concept of a prior. The full distribution of unknown quantities is called the posterior distribution $p(\theta|\mathcal{D})$. The posterior distribution can be expressed as $p(\theta|\mathcal{D})$ by conditioning the latent variable θ on the data $\mathcal{D}$. The construction of the posterior involves Bayes' rule which expresses the posterior as the likelihood of the data $p(\mathcal{D}|\theta)$ multiplied with a specified prior over the latent variables $p(\mathcal{D})$, normalized by the marginalized evidence $\int p(\mathcal{D}, \theta)d\theta$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta)}{\int p(\mathcal{D}, \theta)d\theta} \tag{1.2}$$

Though concise in mathematical notation this model is intractable for any non-trivial application. The main difficulty lies in the marginalization of the evidence term $\int p(\mathcal{D}, \theta)d\theta$. Computing the evidence requires the summation over every possible latent variable configuration. Even for a binary latent variable, the evidence requires 2^N computations, where N is the number of latent variables. The sheer number of computations required for the evidence term makes an exact solution computationally impossible [MacKay, 2003]. The benefit of this approach is that in the Bayesian framework, every relevant question about unknown quantities is framed in terms of the posterior distribution. Therefore a major aspect of Bayesian models is their tractable inference, which amounts to finding scalable approximations to the evidence term $\int p(\mathcal{D}, \theta)d\theta$. Approximation techniques have traditionally focused on different Monte Carlo (MC) methods of integration. Research here has focused on aiding the underlying sampling mechanism, auxiliary constructs such as Hamiltonian Monte Carlo or Markov Chain Monte Carlo ensured that the sampling space is restricted [MacKay, 1998]. However, these methods still struggle with convergence and the ability to scale to very large datasets [Blei et al., 2017]. A recent trend in Machine Learning is to reframe the question of posterior inference as a search problem. First found in Statistical Mechanics to approximate partition functions for exponential distributions has now found its place in Bayesian Machine Learning. Variational Inference (VI) casts the problem of posterior inference as a gradient-based optimization problem. VI avoids the computing the evidence term by using a lower bound on the evidence. With the lower bound and a distance measure for probability density functions, commonly the Kullback-Leibler Divergence (KL-Divergence), the optimization objective Evidence Lower BOund (ELBO) can be constructed. The ELBO can be used to search for the best possible member of a variational family. Variational families are a set of probability distributions. A common example is the mean-field variational family. The main objective of VI is to find the member of the posed variational family that is closest, in terms of KL-divergence, to the true posterior by optimizing the ELBO. VI has been successfully established as a scalable alternative to MC methods [MacKay, 2003].

Unsupervised methods have been widely used but they are of particular interest in topic modeling for natural language. Natural language refers to Human language digitized into a representation and topic modeling refers to the process of finding a well-describing topic for a document. For this thesis, the conceptual unit of language is a document. The length and interpretation of a document depends on the context. A document can refer to a sentence, an article or even a book. Topic modeling is the task of algorithmically finding a set of groups

for documents. The aim of topic models is to find a set of labels in which groups are similar and the documents belong to the same topic, thus the name topic modeling. A common representation of language is the bag of words representation. Bag of words is a discrete representation of the document in which each document is represented as a histogram of words [Murphy, 2013]. Each histogram is a fixed length vector that counts the word occurrences of every word in all given documents and represents them as a natural number. One approach to probabilistic topic modeling is the Latent Dirichlet Allocation (LDA). The model's name suggests two important aspects: Latent and Dirichlet. A Dirichlet distribution is the multivariate generalization of Beta distribution [Murphy, 2013]. The Beta distribution is a form of meta-distribution for probability distributions. The Dirichlet distribution generalizes this notion to multiple probability distributions. Latent is referring to the fact that each observation of the topics is conditioned on observations. Each observation is only the words occurred in a document, the actual topics never appear alone. The topic structure is this latent and can only be gauged from the conditional distribution. LDA falls precisely into this framework. The observed variables are the words of the documents; the hidden variables are the topic structure; and the generative process is as described here. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables are given the documents. LDA assumes an underlying generative process which produced a mixture of different topic distribution over each document. This generative process defines a joint probability distribution over both the observed and hidden random variables. For example, each document exhibits several topics, LDA seeks to answer: to what extent is each topics exhibit and what words are representative of each topic.

1.2 Research Question

The thesis aims to answer three questions:

1. Does the gradient based estimation influence the results of topic models? [Blei et al., 2003] show that perplexity depends on K, does this also hold for batch size?

2. Does the semantic coherence depend on the type of optimization and K? [Hoffman et al., 2013] recommends research in ELBO optimization for probabilistic models

3. Does ABAE produce topics with higher semantic coherence also with gradient based estimation? [He et al., 2017] suggests this but the comparison is between sampling and gradient based.

At the heart of any research stands the ability to compare to other methods. A major question for unsupervised topic modeling is the evaluation of results. While supervised models have a correct annotation to compare to, unsupervised models require different metrics. Traditional literature proposes several clustering evaluation metrics such as B^3 and Rand Index, since each topic forms a cluster of words those can also be used for topic modeling. These metrics require a set of labels to compare to. Semantic Coherence is a clustering evaluation metric tailored for gauging the human level semantic coherence of topics [Mimno et al., 2011].

Other than the previously mentioned B^3 and Rand Index semantic coherence does not require labels to compare to. Semantic coherence focuses on co document frequencies to construct a semantic measure. Other works have compared Bayesian and Frequentist unsupervised learning techniques but this works will only compare gradient based estimated models. Attention-Based Auto-Encoder (ABAE) claims to outperform LDA in terms of semantic coherence. The original paper only compares with a LDA sampling technique, the comparison is this not necessarily on even grounds. Other topic modeling approaches have criticized LDA for their poorly constructed topics [He et al., 2017]. An ABAE is a topic modeling approach which leverages an attention mechanisms. The claim here lies that an attention-based mechanism is able to come up with more coherent topics [He et al., 2017]. This work will investigate that claim.

Not only are the consequences of sampling vs variational inference not well established, but also the choice of optimization technique for variational inference remains an open question [Blei et al., 2017]. The question of how to best optimize the ELBO remains open research. While Frequentist gradient-based methods underwent extensive research in regards to the optimizer used the choice of optimizer for ELBO has been less explored [Blei et al., 2017]. VI optimizes the ELBO which is more complex objective than ML. The question for the second experiment is thus: which gradient-based estimation technique is best suited for optimizing the ELBO of LDA?

1.3 Outlook

This thesis is structured into four chapters. This chapter has outlined the problem and the main research questions. The next chapter will introduce the necessary theoretical foundation for understanding the main two experiments. In the third chapter, this thesis describes the two experimental setups as well as a description of the data set used. This includes a description of the used metrics, data, and processing. After the experiment has been explained the fourth chapter will show and discuss the findings. The last chapter will distill the previous sections and conclude with some finishing remarks as well as ideas for future work.

Chapter 2

Background

This chapter introduces the relevant theory and concepts for the experiments in the following chapters. The experiment will focus on a comparison between LDA and ABAE. To understand the comparison of the experiment the reader is required to understand the following terms: Natural Language Processing, Latent Dirichlet Location, Variational Inference and Attention Based Aspect Extraction. The ensuing background chapter will provide the necessary background for the reader to understand the experiments later in this thesis. First, the general concept of NLP and their subfields topic modeling and aspect extraction are introduced. With the framework given two different models and their estimation techniques are explained.

2.1 Basics

2.1.1 Natural Language Processing

Natural Language Processing is the process of finding and discovering patterns in human language [Goldberg and Hirst, 2017]. While machine learning is the unrestricted search for hidden patterns, NLP is concerned with applying machine learning to documents written in human language. Human language data covers a very broad spectrum of use cases, a common example includes news articles, chat messages, and academic literature. Virtually every document written by in human language classifies as natural language. Tasks in the field of NLP range of parsing, speech recognition to topic modeling and many more. Natural Language is a convoluted area of interest since its highly interdisciplinary and so fundamental to what makes us human. Human language is inherently symbolic, one could argue that the logic underlying computations is ideal to extract symbolic relations. Even though its symbolic and every human can use language it has been historically an elusive endeavor to understand human language with machine learning. Human language is considered a complex topic due to its ambiguity, context-dependency, and high variability [Goldberg and Hirst, 2017]. To conquer these complexities research has been shifting to statistical approaches for understanding language. While initially focused on linear methods such as regression and support vector machine, recent work has utilized the flexibility of non-linear methods [Goldberg and Hirst, 2017]. In particular neural networks and recurrent neural networks have had a profound impact and

replaced many traditional approaches [Goldberg and Hirst, 2017].

An essential building block for Natural Language Processing is the representation of human language document as features. Though all NLP models are concerned with the digitized representation of language there are mainly two approaches to representing a document: discrete and continuous. The discrete or continuous nature of representation presents a tradeoff between simplicity and interpretability. This thesis only inspects discrete representations of language. The most prominent member of language representations is bag of words (BOW). The continuous counterpart to BOW is continuous bag of words (CBOW).

Independent of the representation of language every model will have set of documents and each document will have at least one word. $y_{il} \in 1, ..., V$ represents the identity of the l'th word in document i. In this set V is set of possible words over all documents in the set [Murphy, 2013]. We assume $l = 1 : L_i$, where L_i is the known length of document i and $i = 1 : N$ where N is the number of documents. Within this framework, the BOW representation for topic modeling can be formalized with the following notation. A major drawback of the BOW approach is to neglect the order in which words occur in the document. This approach represents a single document as histogram vector word occurrences for each word in the entire corpora for each document. Since the number of all words used in a static corpus never change and not all words appear in all documents the length of the feature vector is fixed and sparse [Murphy, 2013]. This can be formalized as $n_{iv} \in 0, 1, ..., L_i$ where n_{iv} is the frequency of word v in document i. The number of words is bound by the number of documents, thus $v = 1 : V$. The resulting word count matrix has the shape of $N \times V$ which can grow very large and place constraints on the number of documents or vocabulary used [Murphy, 2013]. Fortunately, documents tend to express only a significantly smaller subset of the entire vocabulary V. The document-term NxV matrix is thus sparse, this invites for compression techniques to reduce the size of the matrix[Murphy, 2013]. With this formalization of the actual task of feature extraction can now be described in the following section. The aim of the following sections is introduce the joint probability models of $p(y_i)$ or $p(n_i)$ using latent variables to capture the correlations between words.

2.1.2 Topic Modelling

Topic modeling is the process of discovering thematic structure within text [Blei, 2012]. The found thematic structure can be used to annotate the documents to better organize and summarize the content. This approach falls under the class of unsupervised machine learning since the text is not labeled initially. The term topic modeling has been coined by bleialc. Topic modeling has received attention due to its ability to find structure in a very large corpus of text without little prior knowledge. With the ever-increasing amount of data, unsupervised approaches such as topic modeling have become a viable tool to bring structure into unstructured data. Though very successful in the text domain Topic modeling can also be used to find latent structures in images, genetics data, social networks and many more [Blei, 2014]. This work focuses on its application on text documents. The Latent Dirichlet Allocation (LDA) is a simple and popular approach used in topic modeling. The main concept of LDA is that very document exhibits a set of different topics. Each topic is a distribution over a fixed vocabulary. The following illustration highlights these concepts in a single document.

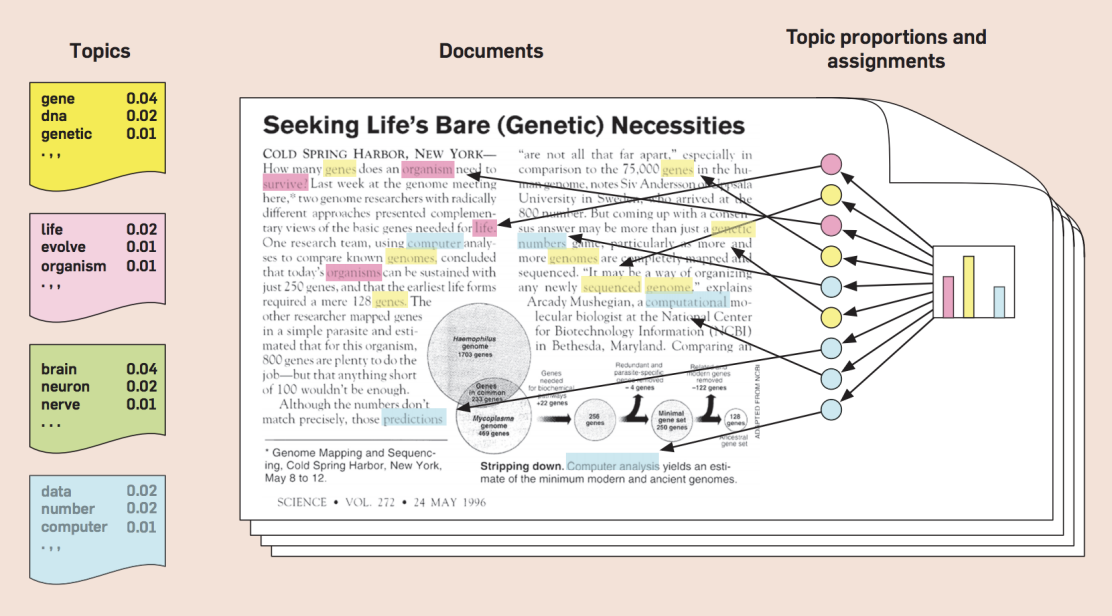

Figure 2.1: A schematic representation of the LDA on article [Blei, 2012]

Figure 2.1 illustrates the concept of topic, document and the topic proportions and assignments. We can see the annotated document contains a color for each word which is associated with a topic. Each document is a mixture of topics, to be specific: each document exhibits a distribution over different topics. This distribution is highlighted in the topic proportions and assignments. The proportion and assignment illustrate the distribution over the colors, ie. the topics. With this approach, one find the distribution over topics for a given document. One can also look at the top words for each topic. The top words serve as a good description of the topics of nature.

The above example highlights that LDA is a useful approach to finding topics within the unlabeled text. The next section will explain the computational problem when using this approach.

2.1.3 Probabilistic Modelling

Finding hidden patterns in data stands at the core of machine learning. One way to gauge hidden patterns is to model them in terms of unobserved random variables that capture patterns in observed data. One approach to topic modelling is to treat the topic distribution as an unobserved random variable. A probabilistic model of a latent variable z and given data x describes the posterior distribution as $p(z|x)$. The construction of the posterior involves Bayes' rule which expresses the posterior as the likelihood of the data $p(x|z)$ multiplied with a prior over the latent variables $p(z)$, normalized by the marginalized evidence $\int p(x,z)dz$

$$p(z|x) = \frac{p(x|z) * p(z)}{\int p(x,z)dz} \tag{2.1}$$

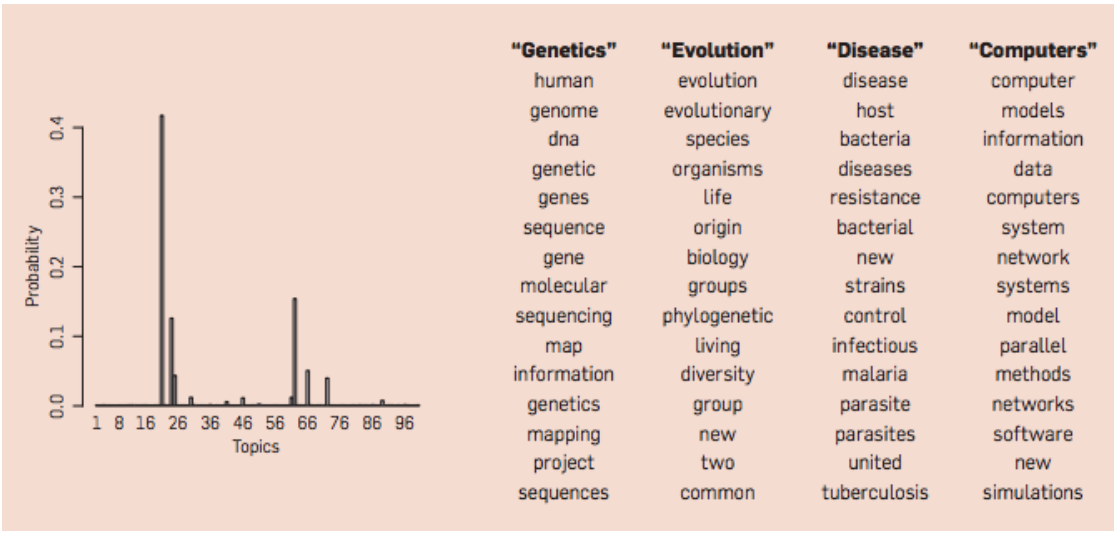

Figure 2.2: On the left are the topic probability for the previous example toping and on the right are the top 15 words for 4 topics [Blei, 2012]

Though concise in mathematical notation this model is intractable for any non-trivial examples. The main problem here lies in the marginalization of the evidence term. Computing the evidence requires the summation of every possible latent variable configuration. For a binary latent variable, the evidence requires 2^N computations, where N is the number of latent variables. There are two general approaches to dealing with the evidence: sampling and variational inference. While sampling such offers conceptual simplicity it struggles with larger models. VI casts the problem of posterior inference an optimization problem [Blei et al., 2017, Wainwright and Jordan, 2008, MacKay, 2003]. One simplified constraint one can employ is to limit the prior to be of the same probability distribution family. Using a conjugate prior allows expressing the posterior in a closed form. This simplifies the problem of posterior inference remains a hard problem as actual inference algorithms make assumptions that will make the integral tough to compute again. The next section will introduce two useful conjugate priors. Conjugate prior!!!

2.2 Latent Dirichlet Allocation

2.2.1 Beta and Dirichlet Distribution

This section will introduce the prior for our probabilistic topic model later on. Since the name of the model inherits from its prior I will commence by explaining its role in the model. The Beta distribution is a probability distribution function (PDF) that has two free parameters: a,b. The Beta distribution is commonly used as a prior in Bayesian modeling since it serves as a conjugate prior for common discrete posterior distributions such as Bernoulli and Binomial. The PDF of the Beta distribution can be given as:

$$Beta(x|a,b) \triangleq \frac{1}{B(x|a,b)} x^{a-1}(1-x)^{b-1} \tag{2.2}$$

The Beta distribution uses the Beta function which is defined by the following equation where is the gamma function. The gamma function is defined for $\Gamma(x) = \int_{\inf}^{0}$

$$B(a,b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a) + \Gamma(b)} \tag{2.3}$$

The Beta distribution serves as a starting block for the Dirichlet distribution. The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector α of positive real numbers. It is the multivariate generalization of the just described Beta distribution. Like the Beta distribution, it is a popular prior in Bayesian modeling. It is used as prior because it is the conjugate prior to the multinomial and categorical posterior distribution. The Dirichlet distribution has the following pdf.

$$Dir(x|a) \triangleq \frac{1}{B(\alpha)} \prod^{K} k = 1 x_k^{\alpha_k^{-1}} I(x \in S_k) \tag{2.4}$$

The I is the mutual information of the data points x in the simplex. The Dirichlet distribution has an interesting geometric interpretation. Dirichlet distribution pdf is a function thas support over the simplex $S_k = x : 0 < x_k < 1, \sum Kk = 1 x_k = 1$.

In section a) of the chart above the full simplex of the parameter vector α is illustrated. This is directly representing the simplex geometry described above. Section b) charts the probability density for the Dirichlet distribution given $\alpha = (2, 2, 2)$, which is broad and centered in the middle. Section c highlights the shape of $\alpha = (20, 2, 2)$ with a tight peak on one corner. d) shows how all small $\alpha = (0.2, 0.2, 0.2)$ values creates peaks on the edges. Small values under 1 create spikes at the corner of the simplex [Murphy, 2013]. This interpretation of the simplex surface will appear in the next section where the words of a document span the simplex.

2.2.2 Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) is one of the main models of the thesis and represents a probabilistic topic modeling approach on discrete data representation [Blei et al., 2003]. The LDA interprets the documents q_i as a discrete mixture model where every document is a mixture of Categorical distributions. The parameters of the Categorical distribution are treated as a random variable and give it a prior distribution defined using the described Dirichlet distribution. Given K as the number of topics, every document is assigned to a single topic $q_i \in 1, ..., K$. The assignment to each topic is drawn from a global distribution π. Every word is assigned to its own topic $q_{il} \in 1, ..., K$ drawn from a document-specific distribution. In this model, every document exhibits membership of every topic to a certain degree or probability. LDA differentiates itself from other approaches where every document only exhibits membership of a particular topic. For this reason, this approach is also referred to as the mixed membership model [Murphy, 2013]. The name Latent Dirichlet Allocation

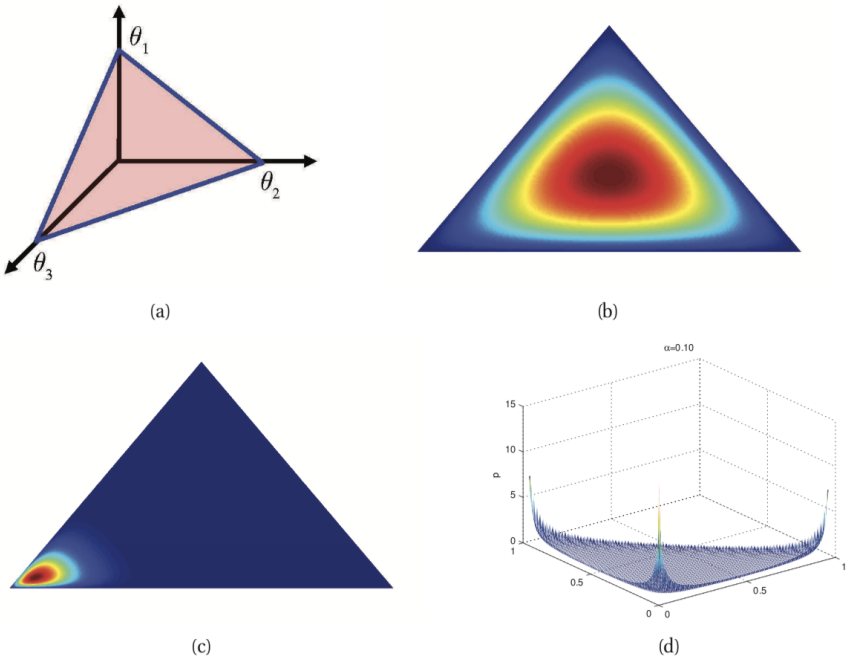

Figure 2.3: The Dirichlet distribution four different parameter values [Murphy, 2013].

comes from the prior used on the categories of document. The Dirichlet distribution is conjugate to the categorical distribution. The relevant posterior is categorical but just the prior on this distribution is Dirichlet. Using a Dirichlet priors makes LDA a twofold process. Not only is the parameter Each parameter for the Categorical distribution Cat() is drawn from Dirichlet prior. With this in mind, the LDA can be formalized as the following. The prior $Dir(\alpha 1_K)$ is specified as the conditional probability of the categories given the values of the prior.

$$\pi_i|\alpha \sim Dir(\alpha 1_K) \tag{2.5}$$

$$q_{il}|\pi_i \sim Cat(\pi_i) \tag{2.6}$$

$$b_k|\gamma \sim Dir(\gamma 1_V) \tag{2.7}$$

$$y_{il}|q_{il} = k, B \sim Cat(b_k) \tag{2.8}$$

The twofold nature of this model becomes in the use of parameters. The distribution over topics for a document is drawn from $q_{il}|\pi \sim Cat(\pi_i)$ which is drawn from $\pi_i|\alpha \sim Dir(\alpha)$.

This hierarchical interaction can also be illustrated in the form of Probabilistic Graphical Model (PGM) [Wainwright and Jordan, 2008]. The advantage of this formalism is the ability to express inference and other interesting functions as the transformation of the underlying graphical model. Ignoring the details of PGM for now, the representation of the LDA as PGM allows to highlights the hierarchical nature of parameters.

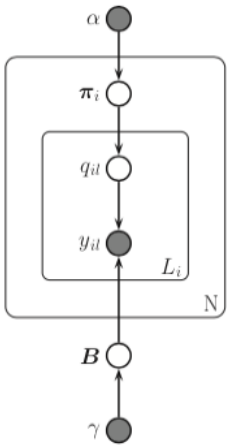

Figure 2.4: Latent Dirichlet Allocation represented as a probabilistic graphical model [Murphy, 2013].

In this graphical notation shaded vertices are observed variables and grey are latent variables. The edges with a direction notate conditional dependencies between variables. The plates refer to repetitions of sampling steps with the variable.

A useful approach to better understand LDA is to understand it from a geometric perspective. As previously described each vector b_k defines a distribution over V words with k topics. Each vector pi_i defines a distribution over K topics.

When only a significantly less number of topics and words are given LDA can also be interpreted as a form of dimensionality reduction [Murphy, 2013]. In this case, topics span a low-dimensional subsimplex and the projection of each document onto the low-dimensional subsimplex can be thought of as dimensionality reduction [Murphy, 2013].

The following figure illustrates this projection with a vocabulary of three (V=3) and two topics (K=2). The resulting V-1 simplex spans a space that represents all probability distribution over the words. The shaded area is the 2-dimensional simplex that represents all possible probability distributions for the three words. In this fashion, documents can be represented as a point in this plane as well. In the resulting three dimensional space observed documents are approximated as being 2-dimension on a 2-dimensional simplex. The 2-dimensional simplex is spanned by the specified two topics, each of which live in a 3-dimensional simplex.

In the illustration above displays the 3-dimensional simplex. The two topics are colored are black, they are represented in terms of the probability for each word. A geometric interpretation of the Dirichlet prior on the topic-word distributions is that it can be interpreted as

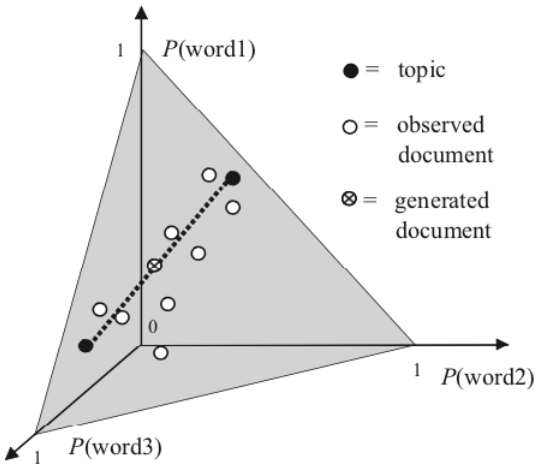

Figure 2.5: Geometric interperetation of LDA with three words and two topics [Murphy, 2013]

forces on the topic locations with higher α moving the topic locations away from the corners of the simplex. This was illustrated in a previous section.

The main question that arises now is: How are the parameters for LDA estimated? This questions stands at the heart of much LDA research. There are two schools of parameter estimation techniques for complex probability density function. Monte Carlo methods, which have a long history of simulating solutions to hard to compute integrals. And the other approach is Variational Inference, instead of approximating the integral this approach solves a similar but easier problem. VI rephrases the problem of parameter estimation as an optimization problem.

2.3 Monte Carlo Estimation Techniques

Computing the posterior distribution for any non-trivial Bayesian model is intractable. In general, computing the distribution of a function of a latent variable using the change of variables formula can be difficult. One simple but powerful approach is to generate S samples from the distribution, call them $x1, ..., xS$. There are many approaches to generate such samples Markov chain Monte Carlo or MCMC is a popular choice. With the generated samples, the distribution of $f(X)$ can be approximated by using the empirical distribution of the generated samples [MacKay, 2003]. Sampling methods are well established but also expose some significant flaws. The intractable nature of the evidence term in probabilistic models has given rise to several sampling methods specialized for machine learning. For the context of this work, two Monte Carlo approaches are relevant: Monte Carlo and Collapsed Gibbs Sampling [MacKay, 1998]. While Monte Carlo is the general approach for approximating complex integral, Collapsed Gibbs sampling is a special use case of sampling commonly used

for LDA.

2.4 Information Theory

Information Theory provides the theoretical underpinnings for scaleable probabilistic topics modeling as well as evaluating the coherence of topics models. At the heart of Information Theory stands the problem of communication. Claude Shannon, the father of this field, has defined the fundamental problem of communication as the ability to reproduce a message that was sent at a different point [MacKay, 2003]. Though the relation to probability and the content of the thesis might not be apparent at first glance it undoubtedly is connected. For example, the distribution over words can be seen as a way to encode a message where common words are bits. A different example would be that the most common words in the English language are very short ('I','and','the') while rare words tend to be much longer [Murphy, 2013]. The relevant part of information theory for this thesis are: Entropy, Kullback-Leibler Divergence and Mutual Information.

The entropy of a random variable X with with a probability density function p is denoted by $H(X)$ or $H(p)$. For discrete random variables with K states the entropy H is defined as:

$$H(X) \triangleq -\sum_{k=1}^{K} p(X=k) log_2 p(X=k) \tag{2.9}$$

The term Entropy comes from the field of Thermodynamics and refers to the disorder of a system. Intuitively higher entropy means more disorder in a system, this disorder interpreted in the context of probabilities as uncertainty. A high entropy implies high uncertainty of outcome. Looking at the Entropy of a Bernoulli random variable highlights this concept. For a Bernoulli distribution with with $K = 2$ and $p(0.5)$ the entropy is at 1. This means that if two outcomes are equally likely the uncertainty is at its maximum. Uniform distributions thus produce high entropy. If the concept of entropy is extended to several probability density functions it can also be used to measure the dissimilarities between random variables. Entropy is this context is called relative entropy and is used to construct the Kullback-Leibler Divergence (KL-divergence). The term $H(p, q)$ is called cross entropy and the extension to the previously described entropy term $H(p)$.

$$H(p, q) \triangleq -\sum_{k} p_k log q_k \tag{2.10}$$

With this definition of cross entropy of two random variables the KL-divergence can be defined as the following:

$$KL(p||q) \triangleq \sum_{k=1}^{K} p_k log \frac{p_k}{q_k} \tag{2.11}$$

$$KL(p||q) = \sum_{k} p_k log\ p_k - \sum_{k} p_k log\ q_k = -H(p) + H(p, q) \tag{2.12}$$

Intuitively the Cross entropy H(p,q) is the average number of bits required to encode data coming from source with distribution p when we use q to define the code book. The code book is referring to the underlying code producing the message. Consequently, the regular entropy is equivalent to the cross-entropy with the same argument. H(p) = H(p,p). Using the frame of information theory on can understand the KL divergence as the average number of additional bits required to encode the data with distribution q instead of the true distribution q [Murphy, 2013]. This measure will play a major role in the following sections as LDA can be estimated by optimizing the KL-divergence.

Several version of LDA will be evaluated in terms of their semantic coherence. Semantic coherence is commonly just referring to Pairwise Mutual information (MI), which is closely related to entropy.

[Murphy, 2013] defines PMI as:

$$PMI(x,y) \triangleq log\frac{p(x,y)}{p(x)p(y)} = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)} \tag{2.13}$$

The PMI measures the discrepancy between two events occurring together versus them occurring together by chance. This can be interpreted as several words occurring together in a topic by the topic model or just by chance. This concept will resurface later as a key metric for evaluating topic models.

2.5 Variational Inference

Variational Inference allows circumventing computing the likelihood of the evidence term by framing posterior estimation an optimization problem. To treat posterior estimation as an optimization problem the objective function needs to be specified and justified. The already introduced KL-divergence is the backbone in constructing a useful objective function for finding a distribution that is similar or close to the true posterior. This section explains how an optimization objective equivalent to the KL-divergence can be constructed. The Evidence Lower Bound is equivalent to KL-divergence but does not require the marginalized evidence term. Furthermore, the Mean-Field approximation as a Variational Family is introduced. This section concludes with alternatives to KL-divergence and the Mean-Field Variational Family.

1: Initialize $\lambda^{(0)}$ randomly.
2: **repeat**
3: **for** each local variational parameter ϕ_{nj} **do**
4: Update ϕ_{nj}, $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.
5: **end for**
6: Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$.
7: **until** the ELBO converges

Figure 2.6: Variational Inference for LDA algorithm [Blei et al., 2003]

2.5.1 The Evidence Lower Bound

There are many approaches to define a measure of distance between two probability density functions. An information theoretic approach is the Kullback-Leibler (KL) Divergence. KL divergence has its origins in information theory and is an asymmetric, nonnegative proximity measure for two densities. Even though KL-Divergence allows to express the distance between two probability function it still requires the computation of the untractable evidence term. The evidence lower bound (ELBO) allows circumventing the intractable computation by optimizing a lower bound on the marginal probability of the observations log p(z). Using the Jensen's inequality, a computational tractable lower bound on log p(x) can be constructed. Jensen's inequality and the concavity of the logarithm function imply that for any random variable y there exists a lower limit for the logarithm of the expectation. The formal definition of the Jensen's inequality

$$log\ \mathbb{E}[f(y)] \geq \mathbb{E}[f(log(y))] \tag{2.14}$$

With this inequality, one can construct a lower bound for the KL-divergence which is denoted as $\mathcal{L}(q)$. The next equation shows that the ELBO and KL-divergence are almost identical and that the ELBO does not require the computation of log p(x):

$$\begin{aligned} log\ p(x) &= log \int p(x,z,\beta) dz d\beta \\ &= log \int p(x,z,\beta) \frac{q(z,\beta)}{q(z,\beta)} dz d\beta \\ &= log \left(\mathbb{E}_q \left[\frac{p(x,z,\beta)}{q(z,\beta)} \right] \right) \\ &= \mathbb{E}_q[log\ p(x,z,\beta) - \mathbb{E}_q[log\ q(z,\beta)] \\ &\triangleq \mathcal{L}(q) \end{aligned} \tag{2.15}$$

The second last line highlights the two components of the ELBO. First the expected log joint $\mathbb{E}_q[log\ p(x,z,\beta)$ and the entropy term of the variational distribution $-\mathbb{E}_q[log\ q(z,\beta)]$. Both depends on the variational distribution over the latent variables $q(z,\beta)$. The ELBO objective is equivalent to the KL divergence up to an additive constant [Hoffman et al., 2013].

$$\begin{aligned} KL(q(z,\beta)\ ||\ p(z,\beta|x)) &= \mathbb{E}_q[log\ q(z,\beta)] - \mathbb{E}_q[log\ q(z,\beta|x)] \\ &= \mathbb{E}_q[log\ q(z,\beta)] - \mathbb{E}_q[log\ q(z,x,\beta)] + log\ p(x) \\ &\triangleq -\mathcal{L}(q) + constant \end{aligned} \tag{2.16}$$

The main advantage of the ELBO is that the KL-divergence can be expressed without the marginal probability of x. As described earlier is, the main problem of probabilistic models is exponential computations required for the evidence term. Circumventing the computation of the marginal likelihood of the evidence allows for a reasonable approximation of many

probabilistic models [Hoffman et al., 2013]. The optimized distribution is then used as a proxy for the true posterior. The solution and the required search for the solution depend on the choice of Variational Family, which role is described in the next subsection.

2.5.2 Mean-Field Variational Familiy

One major assumption in variational inference is the choice of variational family. A family here refers to a set of distributions over the latent variables with its own variational parameters. The aim of the variational family is to restrict the search space by limiting the possible variational distributions. The complexity of variational family directly impacts the complexity of the optimization. More complex families are harder to optimize than simpler families. Mean-Field variational families are a common choice of variational families. The mean-field variational family assumes that latent variables are not correlated and each individually describes by a single factor in the variational density. The power of the mean-field assumption first became apparent in statistical mechanics. The Ising model is used in statistical mechanics to compute interesting properties of a system of magnets. To understand the benefits of the Mean-Field assumption it is beneficial to take a look at the Ising model. The Ising model is a lattice of spins where each spin points either up or down.

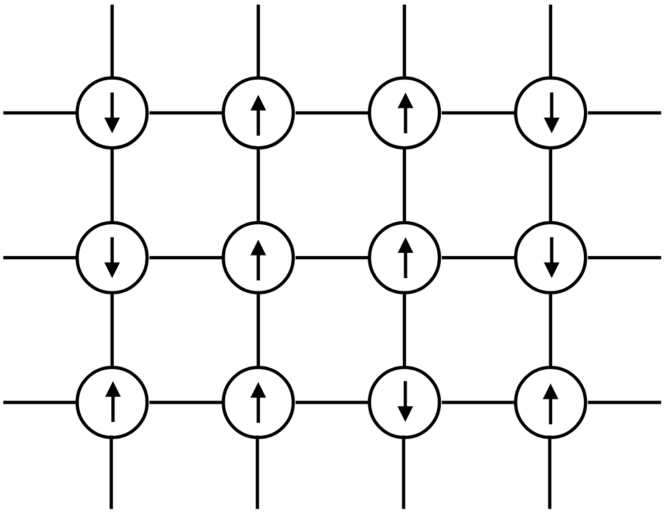

Figure 2.7: Lattice of magnets and their associated spin [Altosar, 2017].

The idea behind the Ising model is that two adjacent magnets can either be attracting or repulsing. Two magnets repulse if their poles oppose and they attract otherwise. The spin at location i is formally described as up $(s_i = +1)$ or down $(s_i = -1)$. The attraction between two magnets is described in terms of the interaction strength J. If two magnets oppose they will contribute $-J$ and J if they attract. For any interaction strength, $J > 0$ the system will align to minimize the energy of the system. The energy of n spins is the sum of all interactions in the system: $E(s_i, ..., s_n)$. Relevant questions such as: what is the likeliest spin or what is the average energy can only be answered in terms of distributions which is

very similar though to probabilistic machine learning [Altosar, 2017]. In statistical mechanics the spin configuration can be described with Boltzmann distribution:

$$p(s_1, ..., s_n) = \frac{e^{-\beta E(s_i,...,s_n)}}{Z} \tag{2.17}$$

The parameter β is the inverse of the temperature and only of anecdotal value in this thesis. The denominator Z is referred to as the partition function Z and is the true highlights of this equation. The partition function Z ensures that the distribution integrates to 1. This is identical to the marginalized probability of the evidence term in probabilistic modeling. Without the right Z the function $p(s_i, ..., s_n)$ would not be a probability. The partition function is given as:

$$Z = \sum_{s_1=\pm 1} ... \sum_{s_n=\pm 1} e^{-\beta E(s_i,...,s_n)} \tag{2.18}$$

This written form the partition function illustrated why this term can't be evaluated analytically. Each s_n has two possible states resulting in 2^n terms for computing the partition function. The partition function is intractable for the same reasons as the evidence term in probabilistic models. However, while in machine learning one is interested in expectations of distribution in statistical mechanics it is the magnetization of the system that is of interest. The solution to the intractable nature of the similar functions was found by Physicists in the seventies [Altosar, 2017]. The Mean-Field Theory is an approximation technique that focuses on certain properties of the distributions such as magnetization and expectations. The Mean-Field approximates the energy by only accounting for local interactions of spins [Altosar, 2017].

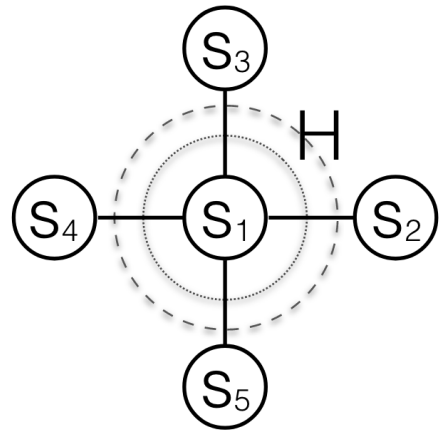

Figure 2.8: The Mean-Field approximation of a single spin. The magnetic field H is shown with dashed lines [Altosar, 2017].

The main assumption here is that the impact of spins on other spins beyond their magnetic field is negligible. One thus only focuses on the z nearest spins. The contribution of each individual spin to the total energy under the Mean-Field approximation can be written as:

$$E_{s1} = -s_1\left(J\sum_{j=1} zs_j + H\right) \tag{2.19}$$

For the two-dimensional case now only needs to sum over the four closest neighbors instead of all other spins in the lattices. Taking this further, the energy for a single spin s_i can also be written in terms of a different spin s_j fluctuation's around its mean value $m = mean(s_j)$ and $s_j = m + (s_j - m)$ The mean-field approximation is now when one even ignores the neighbors fluctuations around its mean and only uses the mean directly. The mean-field approximation for the energy function of a single spin is thus:

$$E_{s1}^{MF} = -s_1(zJm + H) \tag{2.20}$$

Literature refers to this spin as non-interacting since it's energy function only depends on its state s_1 and not of any other states. This example illustrates how the interaction effects of different spins can be approximated by the average magnetic field induced by the neighboring spins. This approach is referred to in Physics as the mean field. Connecting the Ising model to the probabilistic model we can see that the partition function Z is identical to the marginalized likelihood of the evidence term $logp(x)$. Machine learning researched have leveraged the Mean-Field approximation as Variational Family for Variational Inference. The Mean-Field Variational Family describes each latent with its own parameters and no interaction with other latent variables:

$$q(z, \beta) = q(\beta|\lambda)\prod_{n=1}^{N}\prod_{j=1}^{J} q(z_{nj}|\theta_{nj}) \tag{2.21}$$

In this formulation, the term $q(z, \beta)$ is expressed as a member of a mean-field variational family where the posterior now depends on a new parameter λ and θ. One requirement is that $q(z, \beta)$ and $q(\beta|\lambda)$ must be in the same exponential family.

2.5.3 Natural Gradient of the ELBO

One of the most important features of optimization problems is the gradient of the optimization objective. Traditional gradient descents methods optimize a function $f(\lambda)$ by taking steps of size p in the direction of the gradient. The parameter update is described as:

$$\lambda^{t+1} = \lambda^t + p\nabla_\lambda f(\lambda^t) \tag{2.22}$$

If the gradient of function $f(\lambda)$ exists points in the direction of most ascent. Optimizing along the gradient thus guaranties that the change in parameters will improve the objective function. This is formally described as:

$$arg\ max_{d\lambda} f(\lambda + d\lambda) = subject\ to\ ||d\lambda||^2 + \epsilon^2 \tag{2.23}$$

The updates the parameters of our variational distribution are chosen to maximize the negative distance between two distributions. The parameters are updated according to the

gradient of the objective functions. The problem lies here in the assumption that the parameter space is Euclidean. The assumption in 2.25 is that any small change in ϵ would move λ in the direction of the gradient. Euclidean distance is a poor measure for the dissimilarities between two parameter vectors λ and $\lambda^{'}$ and the resulting distribution $q(\beta|\lambda^{'}$ and $q(\beta|\lambda)$. The intuition here is that given a Normal distribution N with two different parameters λ and $\lambda^{'}$ the Euclidean distance between the two parameter vector does not describe the dissimilarities between the distributions. For example, the Euclidean distance between $N(0, 100000)$ and $N(10, 100000)$ is 10 but the actual difference between the two distributions is barely noticeable. On the other hand, the distribution $N(0, 0.01)$ and $N(0.1, 0.01)$ barely overlap but the Euclidean distance is very small. Thus using the Euclidean space for optimizing the gradient would result in poor performance simply because the distribution parameters are assumed to be in Euclidean space. The natural gradient projects the gradient to lie in a space natural to the task. A better approach is to look for the gradient of a space that is similar to the symmetric KL-divergence.

.

$$D_{KL}^{sym}(\lambda, \lambda^{'}) = \mathbb{E}_{\lambda}\left[log\frac{q(\beta|\lambda)}{q(\beta|\lambda^{'})}\right] + \mathbb{E}_{\lambda}^{'}\left[log\frac{q(\beta|\lambda)}{q(\beta|\lambda^{'})}\right] \tag{2.24}$$

The direction of the steepest ascent can also be formulated for the symmetric Kl-divergence.

$$arg\ max_{d\lambda} f(\lambda + d\lambda) = subject\ to\ D_{sym}^{KL}(\lambda, \lambda + d\lambda) + \epsilon \tag{2.25}$$

The difference between the two gradient formulations is that the Euclidean gradient points in the direction of steepest ascent in Euclidean space and the natural gradient points in the direction of steepest ascent in the Riemannian space. In Riemannian space, the distance is defined by KL divergence rather than the L2 norm [Hoffman et al., 2013]. To project the gradient into the Riemann space one can define a linear transformation of λ under which the squared Euclidean distance between λ and nearby vector $\lambda + d\lambda$ is the KL-divergence between $q(\beta|\lambda)$ and $q(\beta|\lambda + d)$. The Riemann metric $G(\lambda)$ is that linear transformation which transforms the Euclidean gradient into the Riemann space.

$$d\lambda^T G(\lambda) d\lambda = D_{sym}^{KL}(\lambda, \lambda + d\lambda) \tag{2.26}$$

2.5.4 Stochastic Gradient Variational Inference

Scalling variational inference to much data? How do update gradient influence result? How to do stochastic?

Using stochastic variational inference allows for a scalable estimation of probabilistic topics modeling, however the resulting topic models are different. [Hoffman et al., 2013] have shown that stochastic approximation to the gradient in variational inference lead to poorer results. The main reason for this result is that stochastic variational inference is guaranteed to converge on a local optimum but approximating the gradient with few data points may lead to a poor results.

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a data point x_i uniformly from the data set.
5: Compute its local variational parameter,

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

6: Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

7: Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}.$$

8: **until** forever

Figure 2.9: Algorithm for variational inference [Hoffman et al., 2013]

2.5.5 Variational Inference beyond KL-divergence and Mean-Field

The two main assumptions in the previous sections where the choice of metric for the distance of the variational family and the choice of variational-family. Though for the context of this thesis only these two choices will be relevant, the alternative should be mentioned as well. The recent spike in interest in probabilistic models has also sparked research in more powerful metrics and variational families. An early example of a variational family beyond mean-field is the Thouless-Anderson Palmer (TAP) equation approach [Zhang et al., 2017]. The TAP equation approach allows relaxing the assumption of total independence by introducing a perturbative corrections to the entropy term. This approach has been successfully been used in Boltzmann machines and Neural Networks [Zhang et al., 2017]. The KL-divergence approach allows for analytically tractable of conditional probability density functions, however, they have been shown to underestimate posterior variance break symmetry when multiple modes are close and is a comparably loose bound [Zhang et al., 2017]. For those three reasons, different divergence measures have been constructed. The α-divergence is a generalization of the ELBO in which the parameter α controls how much mass placement is enforced in the posterior. The α parameter is bound by $\alpha > 0$ and $\alpha \neq 1$

$$D_\alpha^R(p||q) = \frac{1}{1-\alpha}\ log\ \int p(x)^\alpha q(x)^{1-\alpha} dx \tag{2.27}$$

The mass control in the α-divergence regulations how much the posterior the variational distribution is drawn to areas of posterior probability. The small α values force that the mass is spread out over the posterior while a higher α results in zero-forcing of mass. The KL-divergence is a special case of the α-divergence where $\alpha \to 1$. The α-divergence has also been shown to not only to provide a lower bound on the marginal probability but also an upper bound. This has been successfully been leverage in different machine learning application [Zhang et al., 2017]. Beyond the α-divergence there is the f- a divergence which has an even

more general form of:

$$D_f(p||q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{2.28}$$

Where f can be any convex function with $f(1) = 0$, in the case of the KL-divergence this function is the logarithm. With the previous mentioned Jensen's inequality and specialized f function a tighter bound on the marginal likelihood of x can be given. The short section highlighted alternative approaches for variational inference for a detailed discussion please see [Zhang et al., 2017].

One concern raised in research is tightness of the evidence lower bound [**?**]. One would assume that the closer the bound is to the actual evidence term the better the result. However, the authors have shown that tighter ELBO for variational objectives do not always imply a better approximated model. For a full treatment of this please refer to [**?**].

2.6 Attention Based Aspect Extraction

Latent Dirichlet Allocation has been the dominant approach in topic modeling [He et al., 2017]. LDA's success stems from its ability to summarize a given corpus and conceptual simplicity. Both have created a mature ecosystem for estimating these models, making LDA successful and available approach to topic modeling [Pedregosa et al., 2011]. Critics of LDA have argued that LDA-based models may describe a corpus fairly well but the resulting individual topics have poor quality [He et al., 2017]. The poor quality of individual topics manifests in loosely related words in the aspects. The reasons for the low coherence of aspects have been argued to be twofold. Firstly, LDA neglects word co-occurrence statistics which are the primary source of information to preserve topic coherence. The section on coherence describes that co-occurrences are an essential part of measuring the coherence of topics. The second weakness of LDA is its probabilistic nature. LDA requires to estimate the distribution of topics for each word, for applications with small documents this causes significant problems. Documents with few words have been shown to result in poor coherence in the resulting topics [He et al., 2017]. The Attention-based Aspect Extraction (ABAE) is designed to address both of the shortcomings of LDA. The authors of the original paper describe the model as [He et al., 2017]:

> In contrast to LDA-based models, our proposed method explicitly encodes word-occurrence statistics into word embeddings, uses dimension reduction to extract the most important aspects in the review corpus, and uses an attention mechanism to remove irrelevant words to further improve coherence of the aspect

ABAE's aims to improve the coherence of individual topics by using aspect embeddings that take co-occurrences into account. ABAE is designed to estimate a set of aspect embeddings, each aspect can be represented and interpreted by their nearest or representative words in the embedding space [He et al., 2017]. Every word w in the vocabulary has an associated feature vector $e_w \in R^d$ [He et al., 2017]. The word embedding feature vector is designed

to map words that often co-occur to points in the embedding space that are close to each other. This results in space where the distance between two points is the representative of the co-occurrence of words. The embedding features are represented as a word embedding matrix $E \in R^{Vxd}$ where V is the vocabulary size. Each row of this matrix represents a feature vector for each word. The embedding is designed to not only represent the co-occurrences of words but also to represent the associated aspects. Each word is thus embedded in space of co-occurrences with other words and aspects. The require aspect embedding matrix $T \in R^{KxD}$ where K is the predefined number of aspects. The number of aspects is application specific but tends to be significantly smaller than V. The aspect embeddings are necessary to estimate the words for every aspect of the vocabulary V, to further enhance this process the attention mechanism filters aspect words [He et al., 2017]. The general process follows three steps. The input is a list of indexes for words in a document. Each index is then processed in two steps. First all non-aspect words are down-weighted by the attention mechanism

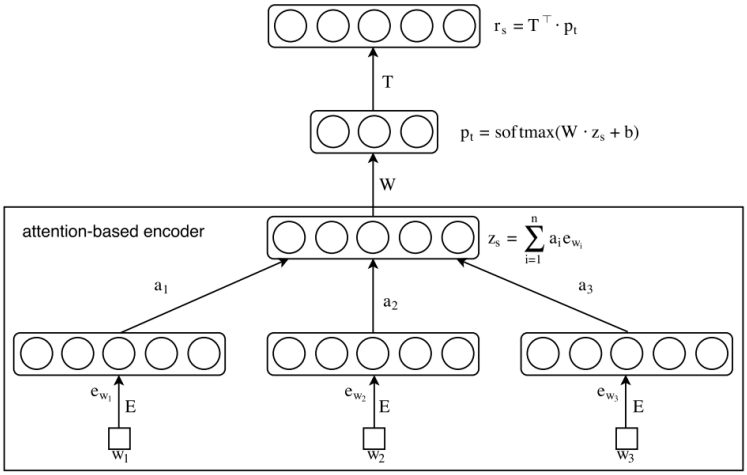

Figure 2.10: A schematic representation of underlying ABAE architecture [He et al., 2017]

Identical to LDA the input to ABAE is a discrete representation of the document, each input sample is a list of indexes for words in a document. The sentence embedding z_s is the weighted sum of all word embeddings e_{wi}

$$z_s = \sum_{i=1}^{n} a_i e_{wi} \tag{2.29}$$

The weight a_i which enforces coherence is estimated by the attention mechanism. The way in which the attention mechanism enforces coherence is by conditioning on the embedding of the word as well the global context of the sentence. The weight can be seen as the probability that w_i is the right word in the order to capture the main topic of the sentence [He et al., 2017]. The weights for a_i need to be estimated with the following procedure.

$$a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)} \tag{2.30}$$

The attention mechanism to compute the weights for a_i follows a two-step logic [He et al., 2017]. First the matrix M captures the relevance of each work to the K topics.

$$d_i = e_{wi}^T \cdot M \cdot y_s \tag{2.31}$$

Then the relevance of the filtered words is expressed as the inner product of the filtered words and the global context y_s. The global context y_s is the average of the words embedding.

$$y_s = \frac{1}{n} \sum_{n}^{i=1} e_{wi} \tag{2.32}$$

The parameters y_s and e_w are estimated by using a gradient-based approach. The authors propose an error function with regularization as well as non-linearities to construct an error function. The error function is then optimized as a neural network. The For a full treatment of ABAE and the associated estimation process see [He et al., 2017].

Chapter 3

Experiment

After the previous chapter has explained the theory behind probabilistic topic modeling and especially LDA with variational inference and ABAE, the coming sections will focus on experiments. The experiments are tailored to address the research questions in Chapter 1. The first section of this chapter describes the data set and the used preprocessing. The following sections explain the used evaluation metrics for the experiments and the actual experiment setups.

3.1 Dataset and preprocessing

The dataset for the experiments come from an online marketplace. Marketplace platforms have become a big part of the economy in the information age. Facilitating a transaction between seller and buyer has not only created economic value but also a glut of transaction data. One vast source of information in those marketplaces are reviews left by users for products. As described in the introduction, the power of unsupervised and especially topics models is to bring structure to vast datasets. Airbnb is a two-sided marketplace platform that allows users to book apartments from other users online. In the year 2017 more than 100 million nights have been booked through the marketplace. After a user has left the accommodation he is asked to write a review for his experience. This includes a rating for cleanliness, location, and communication as well as several written texts. The most salient written review is the public review, the public review is displayed on the apartment's website and visible for future potential guests. The public review is limited to 500 words and the user has 14 days after the checkout to write this review. The public review contains a description of the stay and the overall experience from the user's perspective. For the following experiments, I will use a corpus of Airbnb public reviews. A very similar setup has been used by [Mitcheltree et al., 2018], where the authors focused on ABAE. The corpus contains all reviews written between 1 January 2011 and January 1, 2017. The corpus is then processed with spaCy [1]. SpaCy was used to split the reviews into sentences, remove stop

[1]SpaCy is an open-source library for NLP parsing that leverages Cython to provides an efficient implementation. Cython allows for memory efficient programming in Python. SpaCy is a use full connection between high-level programming in Python and efficient memory allocation in Cython. More information can be found

words and only include the English language. The following two examples highlight how the corpus changes from step 2 to the final step.

```
Tip to share: from Ricky's place to LGA  airport, if you have weekly
metro pass, take sub #1 (8 min walk) to Columbia Univ. and then M60

We called Antonella and she helped us to get to the right address :)

Would definitely recomment to anyone looking for a clean, cozy place to stay.

Thank you so much, Lu!

Great location, comfort and Aaron is a very accomodating person.
```

The above reviews are already split into sentences but still include special characters such as ! , and :). The above sentences also include names and capitalized letters. Both of these characteristics are removed in the next preprocessing steps.

```
train station minute walk away train line station

city center make convenient

place clean nice location

totally grid amazing looking

think hostess accommodating welcoming caring willing make stay truly enjoyable
```

Pre-processing involves removing stop words, removing special characters and normalizing. Spelling mistakes have explicitly not been removed. The full dataset contains five million of the processed sentences. These represent a randomly selected subset of the full dataset. The size should still be large enough to claim that this is a large data set.

3.2 LDA and ABAE Implementation

This section explains what actual implementation of LDA and ABAE are used in the experiments. Scikit-learn is an open source Python library that offers a wide variety of well reviewed and tested machine learning implementations [Pedregosa et al., 2011]. Scikit-learn implements LDA in Numpy and Scipy primitives and provides support for some evaluation metrics. The LDA implementation in Scikit-learn uses Variational Inference based on the work of [Hoffman et al., 2013]. Out of the box, the LDA implementation offers the following parameters[2]:

under https://spacy.io/

[2]The parameter name and descriptions are taken from the official Scikit-learn documentation for LDA http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

```
n_components : int, optional (default=10)

    The number of topics. In the literature, this is referred
    to as K

doc_topic_prior : float, optional (default=None)

    Prior of document topic distribution theta. If the value is None,
    defaults to 1  n_components. In the literature, this is called alpha.

topic_word_prior : float, optional (default=None)

    Prior of topic word distribution beta. If the value is None, defaults to
    n_components. In the literature, this is called eta.

learning_method : 'batch' | 'online', default='online'

    'batch': Batch variational Bayes method. Use all training data in
        each EM update.
        Old 'components_' will be overwritten in each iteration.
    'online': Online variational Bayes method. In each EM update, use
        mini-batch of training data to update the ''components_''
        variable incrementally. The learning rate is controlled by the
        ''learning_decay'' and the ''learning_offset'' parameters.

batch_size : int, optional (default=128)

    Number of documents to use in each EM iteration. Only used in online
    learning where this be used as a degree of noise for the gradient.
```

The LDA implementation in Scikit learn offers many parameter choices for optimization. The relevant parameters for the experiments are "n_topics", "learning_method", "batch_size", "topic_word_prior" and "doc_topic_prior". Since the experiments aim at answering how the gradient update impacts the results of clustering with LDA the parameters not tied to the optimization are ignored. I deliberately focus on only those parameters because the literature and theory suggest a direct impact on the resulting topics. The relationship has been examined in literature [He et al., 2017, Hoffman et al., 2010]. The batch size works as a proxy for noise in the gradient update. The smaller the size the higher the noise for each gradient update. The topic priors α and β are varied for a few experiments since the main focus of this work is gradient updating mechanism.

For the ABAE implementation, the following experiments will use a Pytorch solution described in [Mitcheltree et al., 2018]. The ABAE's underlying optimization problem was solved with ADAM and learning rate of 0.001, the question of the optimizer and a learning

rate of the ABAE implementation will not be touched upon in this thesis. With LDA and ABAE explained the focus shifts to evaluating those two different models. The question of how good a model is can be answered in many different ways. The following section will explain two approaches for evaluating topics models.

3.2.1 Perplexity

One of the classical unsupervised topic modeling evaluation metrics is perplexity. From an information theoretic perspective, the perplexity is a straightforward extension of the introduced cross-entropy term [Murphy, 2013]. Perplexity or hold-out log likelihood is the inverse probability of the test set normalized by the number of words in the vocabulary. Perplexity treats the topic model as a language model, in this context language model refers to the words used in the vocabulary. The perplexity of a language model q given a stochastic process p is defined as [Murphy, 2013]:

$$perplexity(p, q) \triangleq 2^{H(p,q)} \tag{3.1}$$

Where H is the cross-entropy defined in the Information Theory section of this thesis. Perplexity is very similar to plain cross-entropy of the vocabulary and the generated topics. The additions to cross-entropy is simply a different interpretation. Perplexity is interpreted as the weighted average number of choice a random variable has to make. This implies that comparing for two distributions the perplexity gives a weighted average of how many more choice the one distribution has make [Murphy, 2013]. Because of this, perplexity is also referred to as the branching factor. Furthermore, the exponentiation also sets off the logarithm in cross-entropy making the perplexity easier to interpret as the result is in linear space. Low perplexity results are preferred over high perplexity as low perplexity means that the target distribution is has a similar encoding as the original distribution [Murphy, 2013].

A noteworthy drawback to perplexity is, that it has been shown not to correlate with human judgment [Chang et al., 2009]. The authors argue that researchers employ a variety of metrics of model fit, such as perplexity or held-out likelihood, which measures are useful for evaluating the predictive model, but do not address the more explanatory goals of topic modeling. As stated in the introduction of this thesis, topic modeling can be employed to explore topics for a set of texts. Measuring the perplexity does not gauge the semantic coherence of the generated topics. The authors argue that the latent space, the space of topics, behind the model is independent of the perplexity. Other work aims at exactly this, finding the semantic coherence of generated topics [Douven and Meijs, 2007].

3.2.2 Semantic coherence

Besides the classical topic modeling evaluation metrics like perplexity, there is a class of metrics that aim at gauging coherence. As eluded to in the previous section there has been a wide criticism of perplexity as a score for evaluating topic models since they fail to capture the binding power of topics. By the binding power of topics I mean the ability of a topic model, such as LDA, to generate topics that do not only perform well on the test-set but also generate topics where the words have a semantic relation to the other words

in the topic. This is apparent to humans as we can see words in a semantic context and identify of words. Though this is a trivial task for a human it is a hard problem for a topic model. Coherence can be explained as a probabilistic metric that gauges the degree of belonging together [Douven and Meijs, 2007]. The literature emphasizes that not measuring the internal representation of topic models is at odds with their presentation and development [Mimno et al., 2011]. Topic coherence t is represented with the M most probable words for each topic. The m most probable words are arranged in descending order of their collapsed probability. The concept of coherence has been used in many different contexts. The usual choice for a measure of coherence is mutual information.

3.2.3 Pointwise mutual information

The simplest metric for semantic coherence is pointwise mutual information which is the document frequency in relation to the co-document frequency. This formalized as $D(v)$ is the document frequency of word type v and $D(v, v^i)$ is the co-document frequency. The document frequency can be thought off as the number of documents with least one token or work type v. The co-document frequency is the number of documents in which v and $v^{'}$ appears. With this definition, the semantic coherence can be described as [Mimno et al., 2011]:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{3.2}$$

A smoothing addition of $+1$ is added to avoid taking the natural logarithm of 1. Pointwise mutual information or PMI is a common choice for gauging coherence, it is important to note, however, that coherence is always tied into the pairings of words. This implies that the resulting PMI is depended on how the number of words used for word pairings. The word pairings are limited to two for all experiments in this thesis. This is one factor that is consciously neglected though I expect this to have an influence.

3.3 Experiment List

With the full description of the Latent Dirichlet Allocation, Variational Inference, ABAE and evaluation metrics this subsection explains the experiments that aim at exploring the link between gradient updated and the result of topics models. The experiments are tailored to answer the following research questions in more detail.

1. Does the gradient-based estimation influence the results of topic models? [Blei, 2014] show that LDA perplexity depends on the number of cluster K but do not investigate the details of ELBO optimization and influence on perplexity.

2. Since the model's perplexity depends on K as well as the details of ELBO estimation, how does this hold for the semantic coherence of estimated topics? This experiment has a similar scope than the perplexity experiment but this time evaluated in terms of PMI.

3. This question is two-fold. Can the claim of superior semantic coherence of [He et al., 2017] be replicated with a different dataset. Does the higher semantic coherence hold up even when the ELBO is optimized with different algorithms and batch size. Furthermore the role of priors on the results are investigated to ensure that the LDA performance is due to poorly chosen priors.

3.3.1 Gradient update influence on perplexity

The first uses the described dataset to evaluate whether the resulting perplexity. One aspect of the research question of this thesis is to investigate the link between gradient related updates and the result of topic models. The idea behind this experiment is that the link between the number of topics K and perplexity has already been established by [Hoffman et al., 2013]. The authors of the original paper on Variational Inference have use perplexity being sensitive towards K. This experiment goes one step beyond the original analysis and examines whether not only the perplexity is a function of k but also of the number of samples b used in the gradient update. To investigate this question an experiment with the following parameters is specified: number of clusters $k \in 25, 50, 100, 200$ and batch sizes $b \in 10, 50, 100, 500, 1000$. The reasons for this experiment is that results have shown that the perplexity of topics models estimated with stochastic variational inference performs poorer than the using the full gradient [Hoffman et al., 2013].

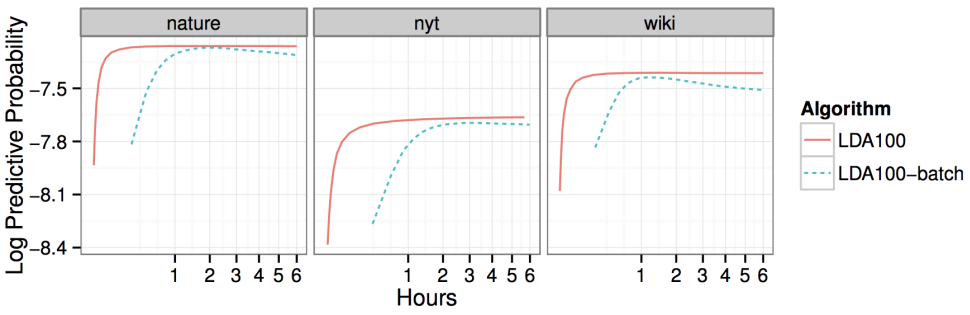

Figure 3.1: Algorithm for variational inference [Hoffman et al., 2013]

The results suggest this intuition as the full gradient outperforms the stochastic approximated one in terms of perplexity. This experiment will be a replication of those findings on a different data set. The dataset and implementation are described in section 3.1 and 3.2. The difference will be that the perplexity will be shown investigated not as a function of time used for approximating the model. The is a replication of the original work but with the addition of gradient update type. With gradient update type I mean the difference between batch and online update to the gradient.

3.3.2 Number of topics influence on semantic coherence

The influence of the number of topics used in LDA and the resulting perplexity has been established by [Hoffman et al., 2013]. The experiment in the previous section is aiming at linking the perplexity to the gradient-based update. Now, this experiment aims at establishing the link between the number of clusters and the semantic coherence of the resulting topics. To examine the link between those two an experiment with $k \in range(2, 150, 5)$. For every k values the model will be computed with batch and online update. As the theory in section 2 suggests they will fit differently purely based on the gradient update. Section 2 explains that the approximation to the gradient is noisy and never fully captures the true gradient. My assumption here is that quality is dependent on the noise introduced by the update. In short, the noisier the gradient update the poorer the resulting semantic coherence. This experiment thus aims at exploring the link between the number of clusters and type of gradient update. To explore this link empirically I will perform the described experiment with the same data of the previous experiment but this time the evaluation metric is the coherence, not the perplexity.

3.3.3 Gradient update influence on semantic coherence

The main question, however, is the influence of the gradient related updates and the resulting semantic coherence of the updates. This experiment goes one step beyond the previous analysis and examines whether not only the perplexity is a function of k but also of the number of samples b used in the gradient update. To investigate this question an experiment with the following parameters is specified: number of clusters $k \in 25, 50, 100, 200$ and batch sizes $b \in 10, 50, 100, 500, 1000$. This setup can be seen as a combination of the previous two experiments. The result of inference in probabilistic models with variational inference is a complex problem, not only is it dependent on model-specific parameters such as the number of topics k but also on optimization specific parameters. This experiment is where the experiment leave the realm of established realms of research as it combines established relations with unestablished relations.

3.3.4 ABAE is superior to LDA in semantic coherence

The authors of the ABAE original paper argued that the addition of an attention mechanism will improve the coherence of generated topics. The authors then proposed an experiment where the semantic coherence of ABAE generated topics is compared to the resulting topics of LocLDA other other prominent topics modeling approaches. The LocLDA is considered as the standard LDA implementation for this experiment. LocLDA uses a Gibbs sampling for estimating the topics. The ABAE experiment used two different datasets to compute topics. (1) The citysearch corpus is a well-established dataset which features over 50000 restaurant reviews from an online platform called city search. (2) BeerAdvocate is a corpus of 1.5 million reviews. Figure 3.2 illustrates the findings of the original ABAE paper.

The findings in the figure 3.2 show that for both datasets the ABAE has a better semantic coherence than the LDA implementation used. In my opinion, the comparison is

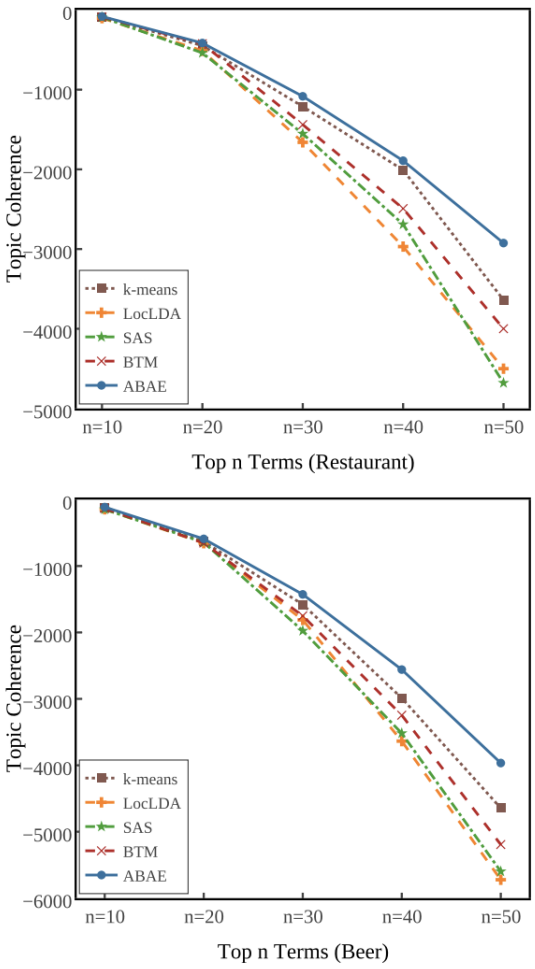

Figure 3.2: The semantic coherence evaluated for ABAE and LocLDA. ABAE consistently outperforms LocLDA [He et al., 2017]

on uneven grounds since the estimation approaches are fundamentally different. The LocLDA implementation uses a sampling approach for parameter estimation while ABAE has been using a gradient-based approach. The differences in result may, therefore, stem from a different estimation approach. The authors of the original paper neglected this aspect in their discussion. Therefore, this thesis will conduct a follow-up experiment that compares a Variational Inference LDA implementation with ABAE. The previous experiments lay the foundations for understanding the relationship between the estimation mechanism and the resulting semantic coherence in generated topics. Given the relationship between this experiment goes one step further by exploring the if different updates to the gradient of the

Variational Inference approach results in a semantic coherence that is better than ABAE. The experiment to test if Variational Inference LDA will be able to have higher semantic coherence than ABAE will use the same sentence review data described for the previous experiments. LDA and ABAE will use the same number of topics 30,60,90,120 but other than the original experiment this experiment will use a Variational Inference LDA implementation. For the first experiment, the stochastic approximation for the gradient will be performed on batch sizes 10,50,100,500,1000. This represents the noise in the gradient update.

3.3.5 ABAE is superior to LDA in semantic coherence with different priors

To make an even fairer comparison the follow-up experiment will take the LDA with the highest semantic coherence and vary the priors α and β. Since the aim is to have a fair comparison this experiment aims trying different priors to ensure not leaving out a major parameters . The role of priors have been explained in the theory section, the figure on the simplex illustrates the influence of different priors. As mentioned in the implementation section, an uninformed prior was used in all experiments up to now. The geometric intuition and heuristics suggests α in $[1.1, 1.0/60, 50.0/K]$ and β in $[0.1, 0.01, 0.001, 1.0/60]$. This experiment aims at giving LDA the change of using the best combination of $K, batch\_size$, α, and β.

Chapter 4

Results

4.1 Gradient update influence on topic model perplexity

This section illustrates and explains the result from the first experiment. The perplexity plotted in figure 4.1 illustrate how the perplexity of LDA estimated with variational inference varies over a different number of topics. The lower the perplexity the better the estimated topics model and in the chart the perplexity appears to increase with a higher number of clusters. These results replicate the findings in where this experiment was performed on a different data set in [Blei et al., 2003, Hoffman et al., 2013]. The same figure also illustrates that the perplexity further depends on the batch size of the optimization algorithm.

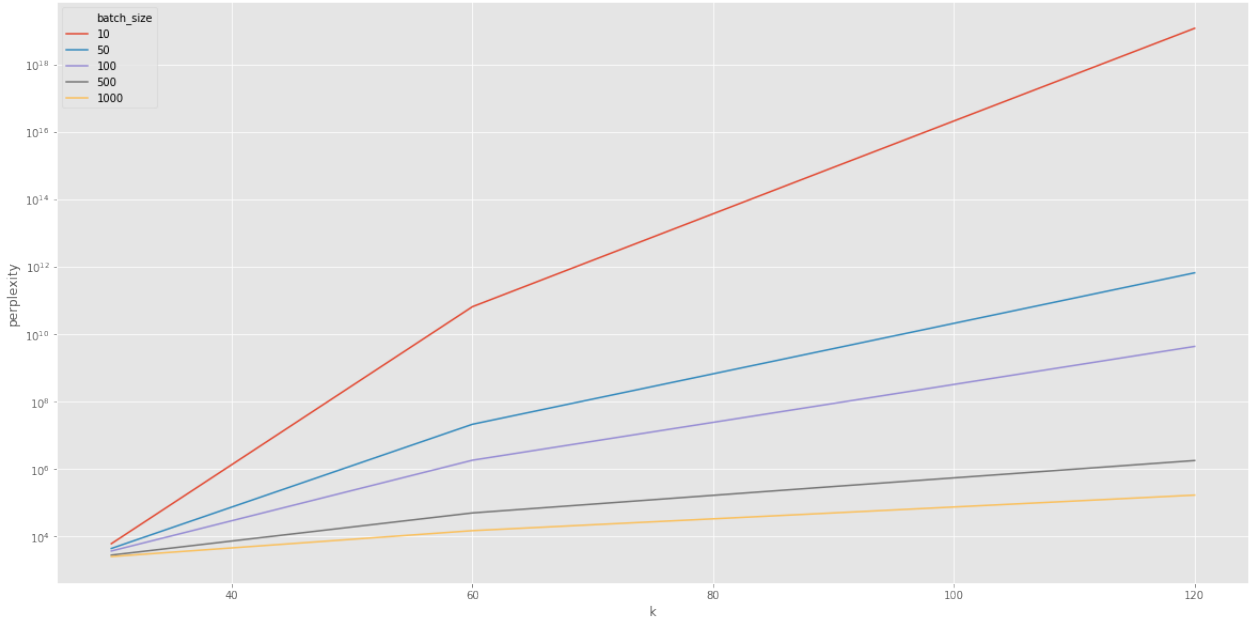

Figure 4.1: Perplexity of LDA models with different gradient update and cluster numbers

The y-axis is plotted on a logarithmic scale because perplexity tends to get very large numbers. The higher number in batch size is more stable against the higher number of clusters. The smallest batch size performs worst and the highest batch size performs best constantly. The results are consistent with previous findings but also introduce the influence on the gradient update on perplexity. The next experiment will repeat the same experiment with coherence as the evaluation metric.

4.2 Gradient update influence on semantic coherence

As the previous results suggest the perplexity depends on the number of clusters and batch size. This experiment illustrates that this relationship also holds for semantic coherence. The higher batch size results consistently in a higher coherence while lower batch size performs inconsistently. The first experiment in this section portraits the impact of the estimation algorithm. This aims at explaining the question raised in [Blei et al., 2017], where the link between gradient related details of ELBO estimation matter for the resulting estimation.

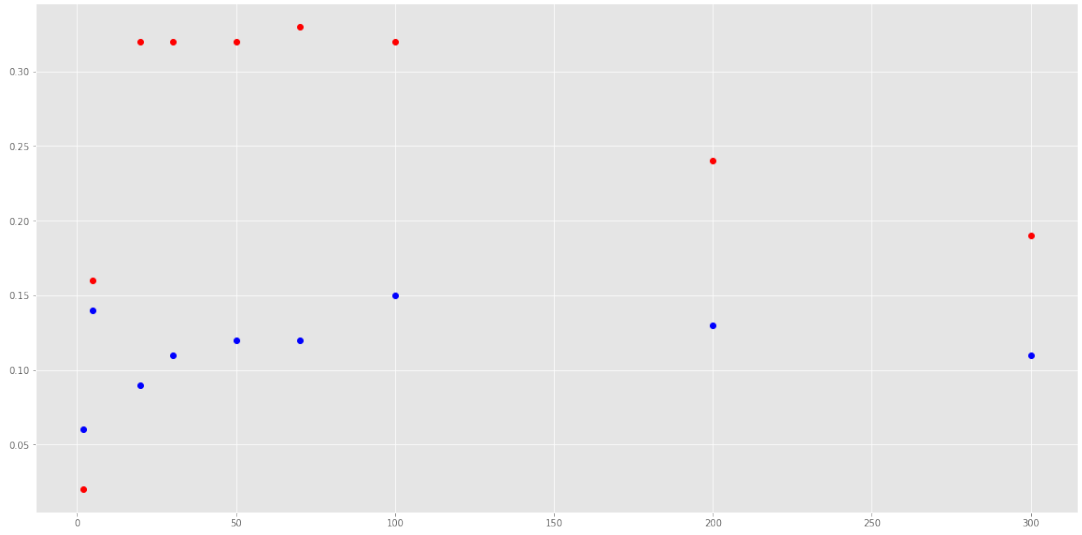

Figure 4.2: Semantic coherence over different k values with batch update and fixed batch in red and online update in blue

Two variational inference LDA models are estimated with different gradient update. This shows that the two perform fundamentally different for different optimizers. Batch update consistently performs better than online update. They both appear to have a peak of semantic coherence around 100 topics and then decrease. For every k values besides 2 the batch update scores higher than online. This suggests that for any relevant model the batch update for the optimization should be preferred.

4.3 ABAE vs LDA

This section explains the results of the replication experiment of [He et al., 2017]. The results indicate the ABAE outperforms LDA in semantic coherence independent of optimization strategy. The LDA results show that the batch size does matter but the best LDA result does not outperform the ABAE results.

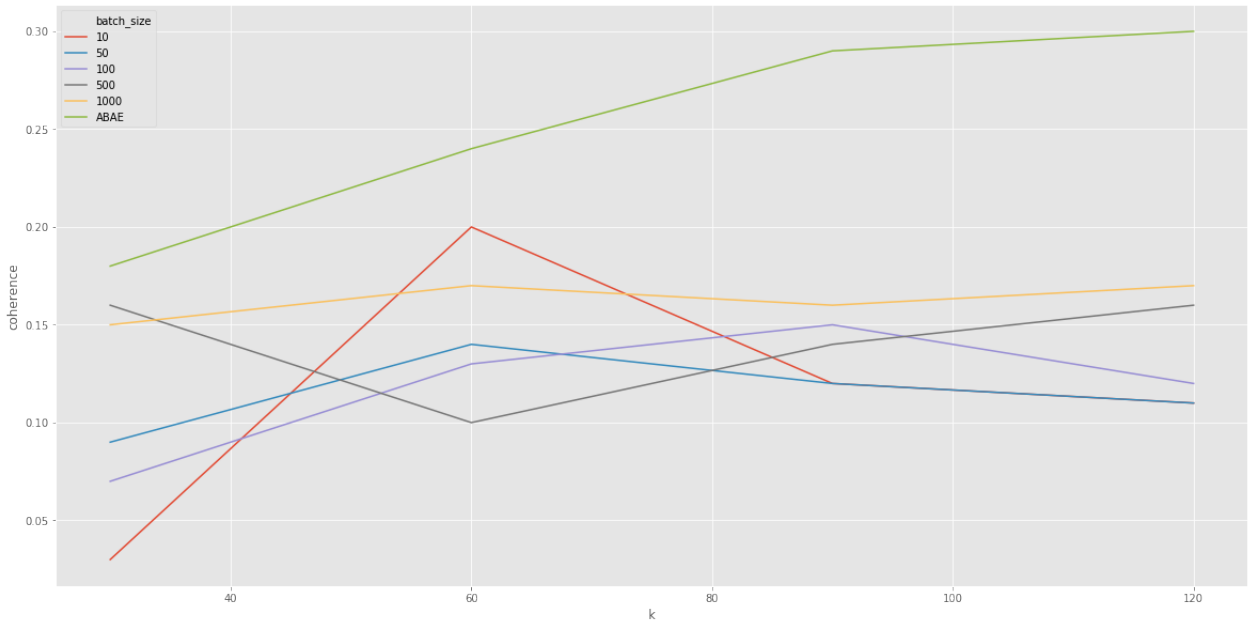

Figure 4.3: ABAE outperforms LDA across K and batch sizes and connected with a line

In figure 4.3 the points where combined with a line to illustrate that the slope and intercept of the underlying relationship. The underlying functional shape shows that ABAE performs better independent of k and batch update details. One can see the ranking of batch size update one the different intercepts of the linear function. We can see that around $k = 60$ the results of the different batch updates are still volatile. This volatility decreases with more K and ends up with a coherence ranking based on batch size. These findings are also in line with the perplexity and batch size relation from the first experiment which adds to evidence that the gradient update of the optimization matters for the result. The batch size represents the noise in the estimator of the stochastic update for the gradient. The more noise the gradient the worse the resulting perplexity and coherence. The batch_size 10 and $k = 60$ presents an exception to general interpretation but this was verified with a follow-up experiment. However, the findings refute my claim that the authors of [He et al., 2017] did an unfair comparison. The introduced coherence enforcing measurements work and produce consistently higher semantic coherence. The following spot checks will highlight the difference between the two. The semantic coherence of ABAE with k = 120 is the highest in the experiment. This is not only apparent in the pointwise mutual information but also for

humans inspecting the top words for a generated topic. The following selection of one LDA and one ABE aspect with the respective highest semantic coherence highlights the difference. Each model has the top 100 words for one topic for the same k values listed below:

- ABAE k=120:
  minimart minimarket delis drugstore 7eleven coffeeshop 24hr bodegas bodega chemist grocer restuarant reataurant pharmacy resteraunt resto brasserie tesco restaurante restraunt cvs restos restaraunt publix supermaket playground starbuck eatery cafes laundromat cafeteria lawson restorant onsite beachside deli pubs nightclub 24hrs bistro retail 24hour coffeehouse restaurants mega eleven laundrette atm hairdresser bookshop lidl shops resturant barber carrefour activity 24h closing 24hours supermarkt poolside patisserie club restaurent shopper hopping closeby lido kiosk takeaway cafs recreation bars parlour gelateria aldi vicinity butcher bank pool boulangerie clubhouse casino beachfront municipal surrounding hawker bookstore resort neighborhood complex arrived spot arcade swimming campground warung creperie coop tiki

- LDA k=120, batch=1000
  close restaurant station near bar metro shop beach cafe several town tourist bike lot surrounding bus major outdoor bakery supply bird connected rock castle link manager great highway scenery shinjuku venue rail cathedral yoga bahn cbd refreshing save penthouse playground kathleen suburb artistic transportation outlet cycling massage deli warmer thai tahoe locally philip pizza cultural tuscany leafy nonetheless champ auckland shin pacific resturants accessibility ancient catarina kreuzberg woken remaining namba rambla lesson shack tavern humour rer bang medina valletta asset rode ebisu walker shoreditch avignon monterey sec avon nation islington backwards citycenter hua dia mick yen liberty spezia akihabara jordaan

Both topics are picked because they appear to me as shopping and food-related aspect. The LDA topics capture a much wider range with terms as "woken", "lesson" and "manager". Words related to actual locations as "kreuzberg" and "jordan" and "medina" only appear in the LDA topics model, this might be a property of the model or just my selection of topic. The yellow highlighted words all refer to the word "restaurant" but most of them are misspelled. ABAE captures this very well and that is likely due to the attention mechanism that enforces co-occurrences of words. This is an interesting side effect has [Murphy, 2013] talks about ABAE's ability to find synonyms but ABAE finds misspelling.

One might argue that the topics of the topic distribution and word topic distribution might play a factor the result of LDA. The LDA is a conjugate prior probabilistic graphical model and the different $alpha$ and $beta$ values will enforce the mass of the different Dirichlet distribution to be close together or spread out. Figure 2.3 illustrates the different mass distributions and Figure 2.5 the geometric interpretation of the prior. The prior has been uniformly distributed on the previous experiment but for this element, they were chosen as described in the experiment section.

Figure 4.4 plots the different $alpha$ prior choice in the y-axis and $beta$ choice in the x-axis, the resulting coherence score is represented with the color. Each coherence score falls into

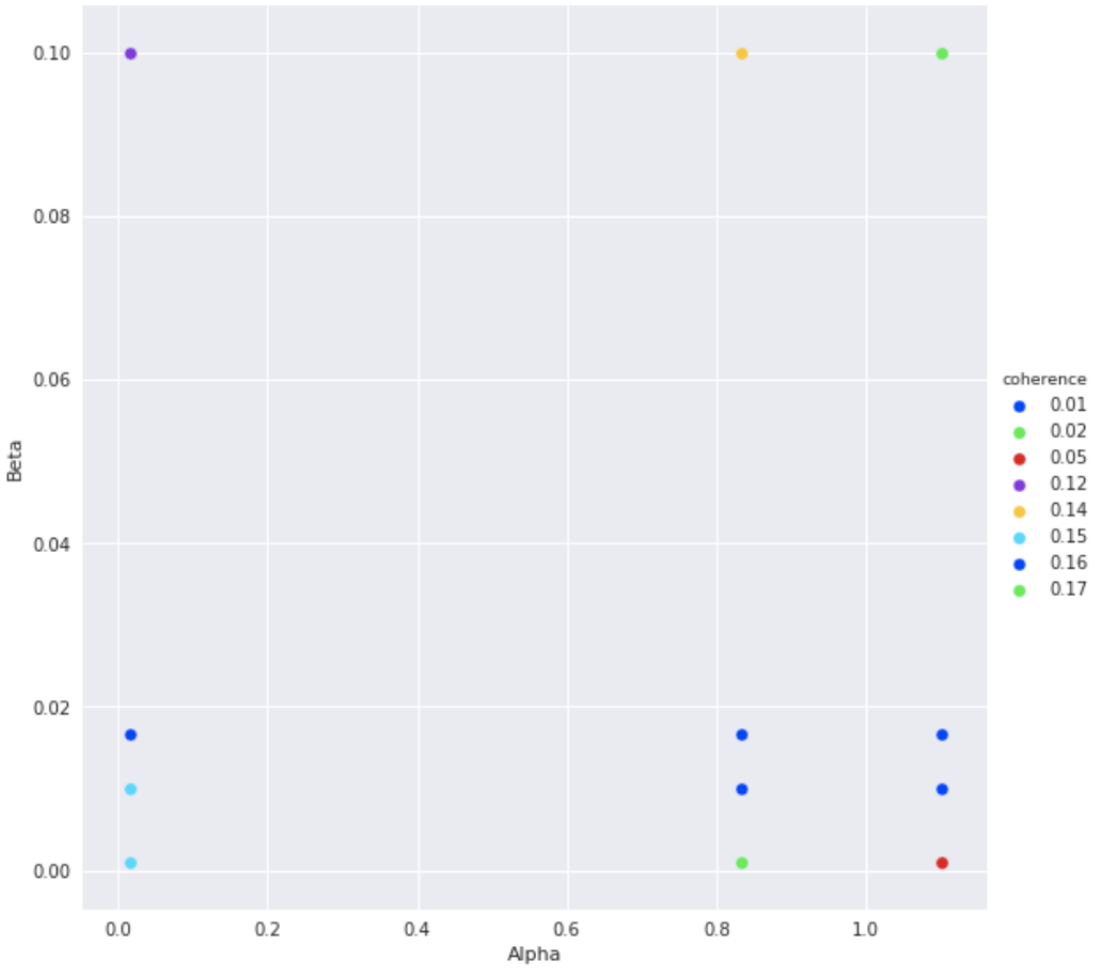

Figure 4.4: The priors (y-scale and x-scale) do not influence the semantic coherence (colours) as directly as the batch update and number of topics

one of the categories listed on the right side of the chart under coherence. It is apparent that there a direct relationship between coherence and both prior parameters. It appears as if the uninformed prior actually works quite well since no change in prior achieves a higher coherence result than the previous experiment. This confirms that ABAE produces fundamentally different results than LDA as ABAE's worst semantic coherence is approximately the same as the best coherence by produced by LDA.

Chapter 5

Conclusion and future work

5.1 Conclusion

This work argues that inference plays a major role in explaining the details of probabilistic models, especially the Latent Dirichlet Allocation. The theory section discussed variational inference along with the ELBO and stochastic variational inference. The experiments of this work aim at replicating [Hoffman et al., 2013] and [He et al., 2017] while conducting further experiments aiming at questions raised in [Blei et al., 2017]. The relationship between the number of clusters and resulting LDA perplexity have been replicated in this work. Furthermore, the link between the number of topics and perplexity has been extended to the semantic coherence of the topics. This work was able to show the link between different EBLO optimizing strategies and resulting perplexity and semantic coherence of LDA. Parameters such batch and online updates produce noticeably different results. This work successfully replicates the findings of [He et al., 2017] where ABAE has received better coherence scores than a sampling-based LDA implementation. The comparison between sampling-based and gradient-based optimization may influence the results. To avoid an unfair comparison this work compared variational inference LDA against ABAE in terms of semantic coherence. The ABAE implementation consistently produces higher semantic coherence in topics independent of estimation technique. The conclusion of [He et al., 2017] therefore still holds that the additional attention mechanism increases the perceived and measured semantic coherence. The attention mechanism of ABAE makes the pointwise mutual information part of the optimization problem. Here the connection between measuring metric and the optimization might be too tight as optimization and evaluation aim at the same quantity. Evaluating both models in term other metrics might yield in a different result as recent work suggests [Chang et al., 2009]. The findings in this thesis showed that optimizing the ELBO a complex problem that requires similar attention as maximum likelihood-estimation and that ABAE consistently produces higher semantic coherence. This work is, therefore, able to contribute to the broader question of topics modeling as an unsupervised machine learning model. The data glut of the 21st century will rely on unsupervised methods and natural language processing to discover latent structure in large data sets. This work showed that not only do the method matter but also the optimization for scaling unsupervised methods to very large

datasets.

5.2 Future Work

After fully discussing the methodology, experiment and results in the last section of this work will discuss future work. This thesis has focused on introducing the intricacies of ELBO estimation of LDA for topic modeling. The attention of this thesis was deliberately set on topic models for natural language. However, the probabilistic machine learning approach is a broad framework can be used for any other machine learning model. The future work is split into three parts.

5.2.1 Variational Inference beyond LDA

LDA has been a useful model to investigate the details of ELBO based optimization for latent variable models. As discussed, the optimization of ELBO is dependent on step size, samples for gradient update and update approach. This thesis established that the result of LDA is dependent on details of optimization processes of ELBO, it is worthwhile to explore this connection in other. Frameworks like Edward or Pymc4 allow solving a broader class of probabilistic machine learning problems with Variational Inference [Tran et al., 2016]. Many other fields also use latent variable models to discover the relation [Gelman et al., 2004]. Since this thesis established that the result of LDA is dependent on details of optimization processes of ELBO, it is worthwhile to explore this connection in other models that use ELBO based optimization. The class of probabilistic graphical models is vast and certainly include candidates for future work. Since LDA is a supervised model, an interesting class would be supervised models and to see if the connection between perplexity and semantic coherence still holds [Wainwright and Jordan, 2008].

5.2.2 Theoretical understanding of the tradeoff in ELBO optimization

This thesis was able to empirically established the connection between the gradient update and resulting semantic coherence of topics and perplexity. Though there exists some theoretical understanding of this is the case there is no full treatment of this problem [Hoffman et al., 2013, Rainforth et al., 2018]. One interesting area of work could be to establish the theoretical framework for finding the ideal step size and batch size for classes of problems. [Blei et al., 2017] calls for similar research and highlights that ELBO optimization requires more research for developing a framework to understand the trade-offs between different gradient optimization techniques.

5.2.3 Other optimizers of ELBO

Beyond the section on alternatives to KL-divergence and Mean-Field Variational Family, there are much more alternatives. There is an emergence in research that aims at reformulating existing gradient based optimization processes such as Root Mean Squared Error (RMSprop) to

optimize ELBO called (VPROP)[Emtiyaz Khan et al., 2017]. Other approaches such as Automatic Differentiation Variational Inference (ADVI) have been implemented in well reviewed frameworks such as Stan [1]. Black Box Variational Inference (BBVI) is a further candidate for an alternative approach to ELBO estimation which has been implemented in BayesFlow [2]. The underlying principle of variational inference remains the same in all these framework, however the details of the gradient based optimization differ greatly. This thesis has shown that small changes to the gradient based estimation processes can have a significant impact on the resulting topics of LDA, changing major assumptions with ADVI and BBVI is certainly a worthwile investigation. The existing implementations invite for revisiting the questions of this thesis but with a different framework to see if the results were due to the Scikit learn approach.

[1] Stan offers full Bayesian statistical inference with MCMC sampling (NUTS, HMC), approximate Bayesian inference with variational inference (ADVI) penalized maximum likelihood estimation with optimization (L-BFGS) and can be found at http://mc-stan.org/

[2] Edward was merged into Tensorflow and is now part of their core API. Details can be found at https://github.com/tensorflow

Bibliography

[Altosar, 2017] Altosar, J. (2017). https://jaan.io/how-does-physics-connect-machine-learning/. Accessed on 05.10.2018.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Communications ACM*, 55(4):77–84.

[Blei, 2014] Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1):203–232.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates.

[Check Hayden, 2015] Check Hayden, E. (2015). Genome researchers raise alarm over big data. *Nature*.

[Douven and Meijs, 2007] Douven, I. and Meijs, W. (2007). Measuring coherence. *Synthese*, 156(3):405–425.

[Emtiyaz Khan et al., 2017] Emtiyaz Khan, M., Liu, Z., Tangkaratt, V., and Gal, Y. (2017). Vprop: Variational Inference using RMSprop. In *Bayesian Deep Learning workshop, NIPS*.

[Gelman et al., 2004] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.

[Goldberg and Hirst, 2017] Goldberg, Y. and Hirst, G. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

[He et al., 2017] He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2017). An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30*, pages 388–397.

[Hoffman et al., 2010] Hoffman, M. D., Blei, D. M., and Bach, F. (2010). Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 856–864, USA. Curran Associates Inc.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal Machine Learning Research*, 14(1):1303–1347.

[MacKay, 2003] MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK.

[MacKay, 1998] MacKay, D. J. C. (1998). Introduction to monte carlo methods. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models*, pages 175–204, Norwell, Massachusetts, USA.

[Mimno et al., 2011] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, Pennsylvania, USA. Association for Computational Linguistics.

[Mitcheltree et al., 2018] Mitcheltree, C., Wharton, V., and Saluja, A. (2018). Using aspect extraction approaches to generate review summaries and user profiles. *North American Chapter of the Association for Computational Linguistics*.

[Murphy, 2013] Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, first edition.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Rainforth et al., 2018] Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4274–4282.

[Tran et al., 2016] Tran, D., Kucukelbir, A., B. Dieng, A., Rudolph, M., Liang, D., and M. Blei, D. (2016). Edward: A library for probabilistic modeling, inference, and criticism.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations Trends Machine Learning*, 1(1-2):1–305.

[Zhang et al., 2017] Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2017). Advances in variational inference. *CoRR*, abs/1711.05597.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Intelligent Adaptive System selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den 16.10.2018 Vorname Nachname

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den 16.10.2018 Vorname Nachname