Pattern Recognition Letters 125 (2019) 310–317

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Boosted negative sampling by quadratically constrained entropy maximization ☆

Taygun Kekeç [a,*], David Mimno [b], David M.J. Tax [a]

[a] Pattern Recognition and Bioinformatics Laboratory, Delft University of Technology, Mekelweg 4, Delft 2628CD, The Netherlands
[b] Information Sciences Department, Cornell University, Ithaca, NY 14853, The United States

ARTICLE INFO

Article history:
Received 31 October 2017
Available online 1 May 2019

Keywords:
Word embeddings
Contrastive learning
Negative sampling
Entropy maximization
Semantic similarity

ABSTRACT

Learning probability densities for natural language representations is a difficult problem because language is inherently sparse and high-dimensional. Negative sampling is a popular and effective way to avoid intractable maximum likelihood problems, but it requires correct specification of the sampling distribution. Previous state of the art methods rely on heuristic distributions that appear to do well in practice. In this work, we define conditions for optimal sampling distributions and demonstrate how to approximate them using *Quadratically Constrained Entropy Maximization* (QCEM). Our analysis shows that state of the art heuristics are restrictive approximations to our proposed framework. To demonstrate the merits of our formulation, we apply QCEM to matching synthetic exponential family distributions and to finding high dimensional word embedding vectors for English. We are able to achieve faster inference on synthetic experiments and improve the correlation on semantic similarity evaluations on the Rare Words dataset by 4.8%.

1. Introduction

The combination of large, publicly available text collections and distributed word vector representations [4] has revolutionized our ability to study the underlying structural patterns of language. Distributed representations, or word embeddings, operationalize the distributional hypothesis [11], which asserts that words acquire meaning over time through their contexts. Embeddings approximate these contextual meanings by mapping words to continuous vectors, so that words that occur in similar contexts have similar vectors.

Recently, studies have shown that these vectors yield substantial representation power and proven to be much more useful in many linguistic tasks than traditional counting based N-Gram representations [3]. Nowadays, word embeddings are typically adopted as fundamental building blocks for a variable set of linguistic tasks [5]. Some successful applications of such vectors are sentiment classification [27], sarcasm detection [12], question answering [6], cross-language text classification [21], and recommendation systems [29].

Word embeddings are typically dense and have radically lower dimensionality than the number of words in a language, but they are nevertheless still high dimensional. Traditional statistical estimators such as Maximum Likelihood Estimation (MLE) easily become intractable for learning these high dimensional models [15]. Negative sampling on the other hand, derived from contrastive learning, easily scales up to large embedding models. Although scalability is an attractive property itself, the user still has to consider design issues to ensure successful learning with negative sampling. Since we have limited data in many practical word embedding problems, it becomes crucial to use a reasonable sampling distribution in order to fit accurate models.

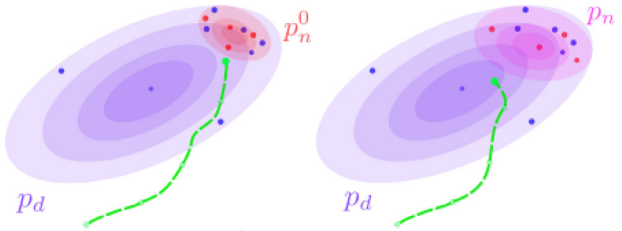
In this work we use negative sampling as the learning component to address aforementioned problems of word embedding architectures. We propose a relaxed Maximum Entropy based sampling principle. The main contributions of this paper can be summarized as follows:

- An objective is obtained which expresses the effect of a sampling distribution with a physical analogy, as attractive and repulsive forces. This formulation lends to a Maximum Entropy formulation.
- A surrogate smoothing objective to the original problem: Quadratically Constrained Entropy Maximization (QCEM) is proposed, posing a computationally attractable method for choosing sampling distributions. Our proofs show that state of the

☆ **Conflict of interest.** None.

* Corresponding author.
E-mail addresses: taygunkekec@gmail.com (T. Kekeç), mimno@cornell.edu (D. Mimno), D.M.J.Tax@tudelft.nl (D.M.J. Tax).

https://doi.org/10.1016/j.patrec.2019.04.027
0167-8655/



(a) Learning with prior p_n^0 (b) Learning with optimized p_n

Fig. 1. Toy example demonstrating the effect of negative sampling distributions on learning. Blue and red points are samples from p_d and the negative distribution. The green trajectory shows the optimization path of the model distribution's mean. (a) Empirical selection of the sampling distribution results in a poor model fit. (b) Optimized sampling distribution pushes away $p_m^{\boldsymbol{\theta}}$ more appropriately and results in a more accurate fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

art heuristics are simple and restricted approximations of our general maximization framework.
- Empirical findings on learning synthetic exponential family densities are provided for analysing the convergence rates of methods.
- The merits of our approach are further demonstrated on word vector space learning when data is scarce and limited. We report word similarity performances on a large number of datasets containing a diverse set of query vocabularies, and find that QCEM-trained vectors had as good or better performance in almost all of the comparisons, and did particularly well on rare words, achieving a 4.8% increase.

2. Quadratically constrained entropy maximization

We are given T i.i.d. data samples $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ drawn from a true but unknown data density $p_d(\boldsymbol{u})$ defined on the real domain $\boldsymbol{u}$. Similarly, negative samples $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ are drawn from the prior negative distribution $p_n^0(\boldsymbol{u})$. The goal is to fit a probability model $p_m^{\boldsymbol{\theta}}(\boldsymbol{u})$, having parameters $\boldsymbol{\theta}$. Without loss of generality of the framework, one can also learn unnormalized models which $\ln p_m^{\boldsymbol{\theta}}(\boldsymbol{u}) = \ln \tilde{p}_m^{\boldsymbol{\theta}}(\boldsymbol{u}) + \mathcal{Z}$, where $\tilde{p}_m^{\boldsymbol{\theta}}(\boldsymbol{u})$ represents the unnormalized density, and $\mathcal{Z}$ is the normalization factor to be learned. Then, the full parameter set to learn is $\{\boldsymbol{\theta}, \mathcal{Z}\}$. This leads to the negative sampling objective:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{p_d}\left[\ln \sigma(\boldsymbol{x}; \boldsymbol{\theta})\right] + \mathbb{E}_{p_n^0}\left[\ln(1 - \sigma(\boldsymbol{y}; \boldsymbol{\theta}))\right] \tag{1}$$

Negative sampling is an instantiation of the contrastive framework. If we had unlimited data, for any sampling distribution, estimation error would be asymptotically normally distributed [13]. However, we are more interested in the word embedding problems where samples are usually considered to be insufficient for learning high-dimensional model densities. In such settings, our samples are finite, and biased.[1] If we have an unsuitable prior $p_n^0(\boldsymbol{u})$, the learned model $p_m^{\boldsymbol{\theta}}(\boldsymbol{u})$ can easily be inaccurate. For illustrative purposes, consider a toy scenario in Fig. 1a where optimization is in the $\mathcal{R}^2$ space. Here, empirical samples obtained from p_d are highly biased and a naive negative sampling prior $p_n^0(\boldsymbol{u})$ is chosen for learning the model $p_m^{\boldsymbol{\theta}}(\boldsymbol{u})$. Negative sampling can not provide sufficient repulsion to stop $p_m^{\boldsymbol{\theta}}(\boldsymbol{u})$ from overfitting to the empirical samples. Instead, given a criterion to optimize the sampling distribution p_n, we could prevent inaccurate model fits as in Fig. 1b. This motivates one to optimize p_n before we perform stochastic updates to the embedding model.

Although Eq. (1) is the standard formulation of the negative sampling, we want to reformulate it to give us an intuitive understanding on the role of the negative distribution. To make the dependency on p_n explicit, we apply mechanical steps (provided in Supplementary Material) and rewrite Eq. (1) jointly in terms of the embedding parameters $\boldsymbol{\theta}$ and the sampling distribution p_n:

$$\begin{aligned} J(\boldsymbol{\theta}, p_n) = {} & \mathbb{E}_{p_d}[\ln p_m^{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{p_d}[\ln(p_m^{\boldsymbol{\theta}}(\boldsymbol{x}) + p_n(\boldsymbol{x}))] \\ & - \mathbb{E}_{p_n^0(y)}[\ln(p_m^{\boldsymbol{\theta}}(\boldsymbol{y})) + p_n(\boldsymbol{y}))] + \mathbb{E}_{p_n^0(\boldsymbol{y})}[\ln p_n(\boldsymbol{y})], \end{aligned} \tag{2}$$

where we have four terms guiding the optimization of model distribution. With this reformulation, we can express the terms using a physical analogy, as attractive and repulsive forces. The first term is the fit term where we require the $p_m^{\boldsymbol{\theta}}(\boldsymbol{x})$ be similar to $p_d(\boldsymbol{x})$. In the second and third terms, the *mixture distribution* of $p_m^{\boldsymbol{\theta}}(\boldsymbol{u}) + p_n(\boldsymbol{u})$[2] is evaluated under the expectation of $p_d(\boldsymbol{u})$ and $p_n^0(\boldsymbol{u})$. This means, this mixture is repulsed to fit to these distributions and can be interpreted as terms to provide regularization to the learning of $p_m^{\boldsymbol{\theta}}$. We denote the second term as the *data repulsion* force, and the third term as the *prior repulsion* for the mixture distribution. If we analyze a single update on $\boldsymbol{\theta}$, model parameters, the fourth term becomes a constant. We can then illustrate in Fig. 2 how the combination of three terms drives the optimization of the mixture distribution.

If we have not sampled any negatives from the prior $p_n^0(\boldsymbol{u})$, then the third and fourth terms do not contribute to Eq. (2). In this scenario, the data repulsion term is the one preventing overfitting to the data samples. When we know that there is strong bias while sampling the data points, we have to learn p_n such that it provides sufficient data repulsion for the mixture $p_m^{\boldsymbol{\theta}}(\boldsymbol{u}) + p_n(\boldsymbol{u})$. This means we want to maximize the data repulsion $\mathbb{E}_{p_d}[\ln(p_m^{\boldsymbol{\theta}}(\boldsymbol{x}) + p_n(\boldsymbol{x}))]$ term for p_n. This is troublesome at first sight, since it looks difficult to disentangle the p_n function. Luckily, two design considerations in word embeddings allow us to bypass this problem.

First, in many word embedding objectives, including Word2Vec [20] and GLoVe [23] embeddings, optimization is done on sufficiently high dimensional spaces, and model parameters are initialized randomly on the space [17]. Under this condition, we can assume that the model likelihood $p_m^{\boldsymbol{\theta}}(\boldsymbol{x})$ for any given sample will be negligibly low right after the initialization. Furthermore, p_n is usually constructed from the empirical distribution which means $p_n(\boldsymbol{x})$ is going to be the dominant term inside the mixture. These two common design practices allow us to instead optimize an upper bound. For any given data point $\boldsymbol{x}$, we consider $p_m^{\boldsymbol{\theta}}(\boldsymbol{x})$ as a constant and inferior quantity and write the upper bound as:

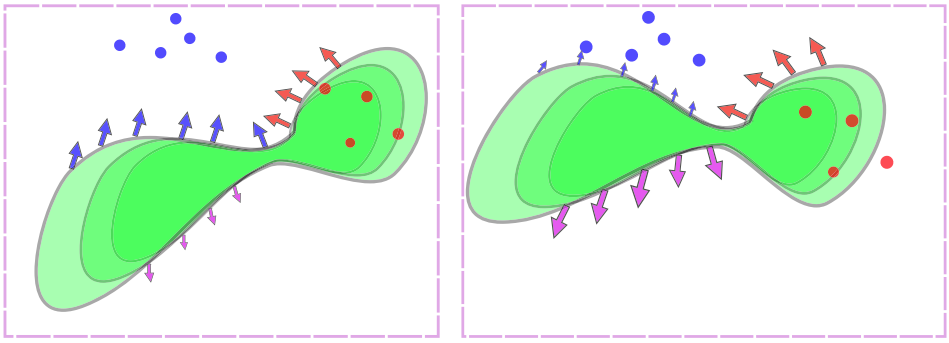
$$J(p_n) = -\mathbb{E}_{p_d}[\ln p_n(\boldsymbol{x})] \tag{3}$$

Since we are trying to maximize the objective, this equation suggests that we want to learn a p_n such that we want to deviate away from the empirical data distribution. The equation is underdetermined in nature; many choices are possible for selecting the a negative sampling distribution p_n. We make the least possible assumptions, and resort to a Maximum Entropy [30] approach given that we satisfy distributional consistency. That is, we want to maximize the entropy of p_n, while being consistent with the empirical data's statistics. Optimizing the upper bound of the data repulsion term with respect to the empirical statistics, we aim to obtain a better sampling distribution for learning $p_m^{\boldsymbol{\theta}}$ as in Fig. 1b. Reliance on the data's empirical moments will constrain the solution.

Assume the initial word frequencies are given in a data vector $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)^T$ in which the entries are ordered: $d_i \geq d_{i+1}$ and where n is the vocabulary size. Let $\boldsymbol{p}$ be the parameters to be optimized for the p_n. We constrain the deviation of $\boldsymbol{p}$ from the data $\boldsymbol{d}$ by a quadratic constraint $(\boldsymbol{p} - \boldsymbol{d})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{p} - \boldsymbol{d}) \leq \beta n$ where $\boldsymbol{\Sigma}^{-1}$ is the precision matrix. These design considerations yield the

[1] Many cooccurrence statistics over word context pairs are either underestimated or overestimated.

[2] Mixture normalization constant is 2 but not shown for the ease of notation.



(a) Early Optimization (b) Late Optimization

Fig. 2. Three forces guiding the optimization of mixture distribution $p_m^{\theta}(\boldsymbol{x}) + p_n(\boldsymbol{x})$, shown in green contours. Blue and red are data and negative samples. Blue arrows represent the fit force, purple arrows represent data repulsion. Red arrows represent the prior repulsion. (a) In early stages of optimization, fit force and prior repulsion push the mixture towards empirical samples. (b) In later stages, data repulsion prevents overfitting to the data. Our goal here is to also optimize the data repulsion term to prevent overfitting to the data samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

following problem:

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \mathbb{H}[\boldsymbol{p}] \\ \text{s.t.} \quad & \boldsymbol{p} \geq 0 \\ & \mathbf{1}^T \boldsymbol{p} = 1 \\ & (\boldsymbol{p} - \boldsymbol{d})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{p} - \boldsymbol{d}) \leq \beta n, \end{aligned} \tag{4}$$

where positivity and sum to one ensures that p_n is a probability function. Although this problem seeks sampling distributions with higher entropy, it is difficult to solve in practice via gradient descent updates. It frequently suffers from numerical difficulties when many probabilities are almost zero.[3] Then, the log function easily yields $-\infty$ values causing the gradient to go infinite where Lipschitz continuity conditions do not hold anymore. As the problem dimensionality increases, we are much more likely to encounter such problems. To circumvent problems arising from entropy maximization, we further want to design a surrogate for the problem in Eq. (4).

Proposition 1. *Let a probability mass function $\boldsymbol{p}$ defined with ordered probability masses: $p_1 \geq p_2 \geq \cdots \geq p_n > 0$. Then the application of a smoothing operator increases the entropy of $\boldsymbol{p}$.*

Proof. The key part of the proof uses a Taylor series expansion. The full proof is provided in Supplementary Material. □

This result poses that there is a relation between the entropy and the smoothing operator. Motivated by it, we relax the entropy maximization problem in Eq. (4) to:

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & -\|(\boldsymbol{\Omega} - \boldsymbol{I})\boldsymbol{p}\|_2 \\ \text{s.t.} \quad & \boldsymbol{p} \geq 0 \\ & \mathbf{1}^T \boldsymbol{p} = 1 \\ & (\boldsymbol{p} - \boldsymbol{d})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{p} - \boldsymbol{d}) \leq \beta n \end{aligned} \tag{5}$$

where $\boldsymbol{\Omega}$ is chosen as a Hankel matrix [9]. This formulation enforces that neighboring entries in $\boldsymbol{p}$ become similar, making the distribution smooth and thereby increasing the entropy. Moreover, the problem is convex in $\boldsymbol{p}$ and known to yield a unique maximum [24]. This formulation does not make any distributional assumption on the form of p_n, nevertheless we can still favour particular solutions by setting the precision matrix $\boldsymbol{\Sigma}$. Using a Hankel matrix in its most general form results in an impractical number of objective terms for large vocabularies. Thus, we further embed a binary structure with $\boldsymbol{\Omega}_{ij} = 1$ if $j = i + 1$ and $\boldsymbol{\Omega}_{ij} = 0$ elsewhere.[4] This specialized circulant structure of $\boldsymbol{\Omega}$ reduces the number of terms in the objective to n, the vocabulary size.

Proposition 2. *Let a PMF $\boldsymbol{p}$ given with ordered masses: $p_1 \geq p_2 \geq \cdots \geq p_n > 0$. Also let $0 < \lambda < 1$ be the density powering parameter. Then, application of powering acts as a smoother on the density given that there exists a lower bound γ on p_i that it is related to λ with: $\gamma = (\frac{1}{\lambda} \sum_j p_j^{\lambda})^{1/(\lambda - 1)}$*

Proof. The proof follows by recognizing the Lipschitz condition, enforcing it to hold by assuming a lower bound and exploiting the diminishing structure of the first order derivative. The full proof is provided in Supplementary Material. □

This result sheds light on why the heuristics [20,23] adopted for negative sampling work moderately well in practice. As long as the minimum probability mass of the sampling distribution is bounded, powering distributions acts as a smoother. This is simply an approximation to our smoothing formulation.

Despite its practical consequences, the problem with the powering heuristic is that, to the best of our knowledge, there is no rationale for the optimal sampling distribution to be in the Pareto family. Unlike [20], which constrains the word frequency density to be in the Pareto family, the formulation in Eq. (5) yields more generality. It does not enforce any distributional assumptions, opening up possibilities to discover better optima. In the next section, we experimentally compare these heuristic approaches to our formulation.

3. Experiments

Experimental setup. We provide two sets of experiments to demonstrate the efficiency of our approach. For both experiments, the QCEM formulation is solved by a Splitting Conic Solver [22] that can solve large-scale convex cone programs by using an alternating directions method [7]. For synthetic experiments, we use inverse transform sampling to sample from 1D probability distributions. The error for model fits is measured by calculating the average $KL(p_d || p_m^{\theta})$ by repeating the experiments 10 times with different random initializations. Learning the model distribution on both synthetic and real world experiments, we use the same

[3] We know that word-context conditional distributions are highly sparse and contain very minor probabilities in their tail.

[4] As Hankel and Toeplitz matrices are closely related, one can question the effect of $\boldsymbol{\Omega}$ being Toeplitz when using this binary structure. In this case, we achieve the same objective with Eq. (5) given that entries of $\boldsymbol{p}$ are reversed. Hence for penalizing the difference of consecutive entries, choosing between Hankel and Toeplitz matrices does not constitute a key difference in our formulation, and is a matter of reparametrization.

stochastic gradient algorithm with the same learning rate for all settings.

3.1. Exponential family density estimation

Data generation and parameters. The interest in this section is to quantify the contribution of QCEM contrastivity for the unnormalized density estimation problem. We define a data generator signal $S(\boldsymbol{\theta}^*, \boldsymbol{\phi}(u))$ over the domain $[-2\pi, 2\pi]$ with sine and cosine bases:

$$S(\boldsymbol{\theta}^*, \boldsymbol{\phi}(u)) = \theta_1^* \sin(2\pi\omega_1 u) + \theta_2^* \cos(2\pi\omega_2 u) + \cdots + \theta_{2n}^* \cos(2\pi\omega_{2n} u),$$

where $\boldsymbol{\phi}(.)$ represents the transformation to the trigonometric functions. Then the probability densities are constructed using the Exponential Family (EF) representation:

$$p_d(u; \boldsymbol{\theta}^*) \sim \exp\left(S(\boldsymbol{\theta}^*, \phi(u))\right),$$

where trigonometric bases are interpreted as sufficient statistics. Finally, we learn the unnormalized EF density $\ln p_m^{\boldsymbol{\theta}}(u) = \ln \tilde{p}_m(u; \boldsymbol{\theta}) + \mathcal{Z}$ with parameters $\{\boldsymbol{\theta}, \mathcal{Z}\}$. In other words, the goal is to learn the true canonical parameters of p_d, the amplitudes of each trigonometric statistic, plus the normalization constant of the density.

Methods. Our first baseline for the contrastive density is the univariate Gaussian density (UNV). Although it is simple to draw samples from this distribution, it is a poor choice for a contrastive function because it is only able to provide a limited amount of discrimination between data and contrastive densities. Another baseline choice of p_c is a more flexible nonparametric kernel density estimate [26] (KDE), where p_c is fitted to the observations. In some applications, one might know the parametric family of the underlying data density in advance, but not its parameters. We depict this case with an Exponential Family (EF) baseline where we have access to the true sufficient statistics of p_d, but not the canonicals. Knowing the true sufficient statistics of p_d is a very strong assumption, making this baseline very competitive. As the synthetic experiments have relatively low numerical complexity, we also report baseline results for the ENT baseline (solution of the Eq. (4)). Finally, QCEM corresponds to our approach with an isotropic precision. We constructed the data constraint vector $\boldsymbol{d}$ for this problem using the Kernel Density Estimate.

Results. Fig. 3 shows the density fits obtained with each negative sampling approach. We observe that the univariate approach can only learn the prominent peaks of p_d in locations with many samples. For instance, the data peaks on the leftmost region are not captured accurately. In contrast, EF collects samples more homogeneously with its trigonometric bases and helps to fit more accurate models compared to the Univariate approach. KDE also obtains a fit that is comparable with the EF and QCEM fits. Using KDE, the low probability region variations are captured, but the probabilities of data peaks are not correctly estimated. QCEM contrastivity obtains the best fits: not only the data peaks, but also the low probability regions are captured much more accurately compared to the KDE and EF.

Note that the distribution p_c obtained by QCEM is relatively uniform compared to the EF and KDE distributions. This might raise the argument that a naive uniform distribution would

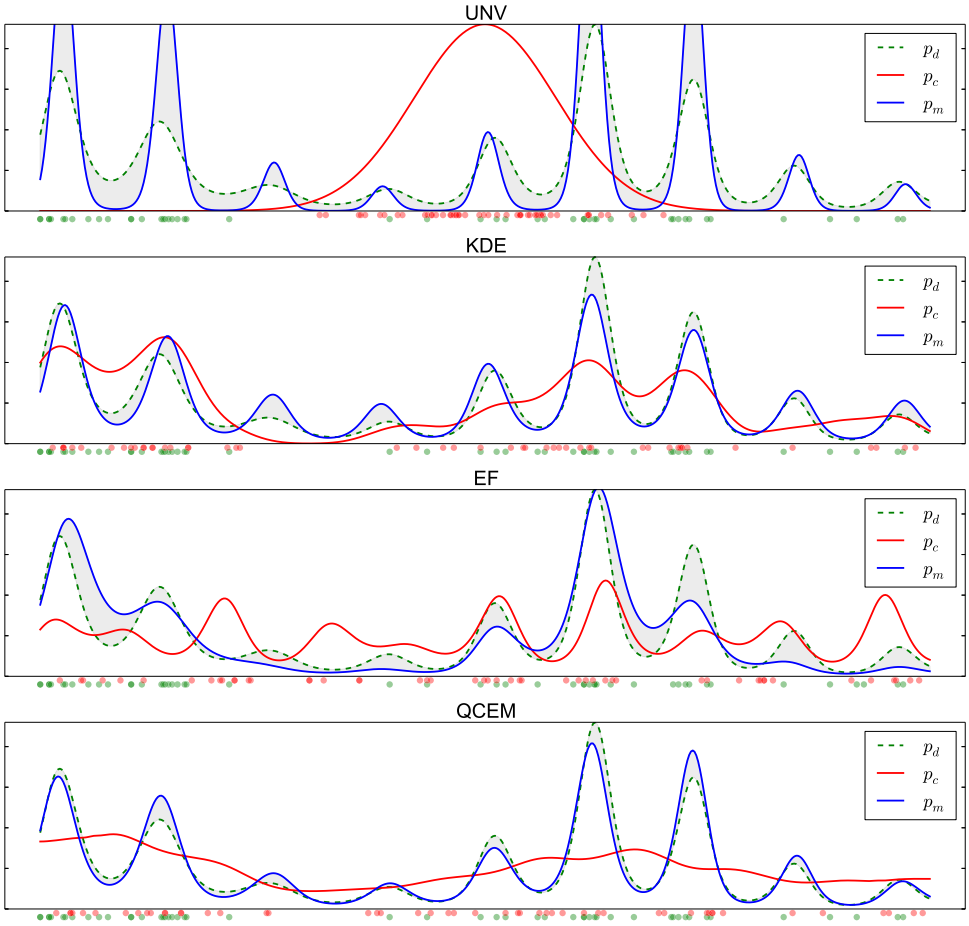


Fig. 3. Learned models (blue) for the data density (green dashed), using different sampling distributions p_c (red). Top to bottom row shows (a) Univariate (b) KDE (c) EF (d) QCEM distributions. Green points are data samples $\boldsymbol{x}$ and red points are negative samples from p_c. Gray areas highlight the fitting errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

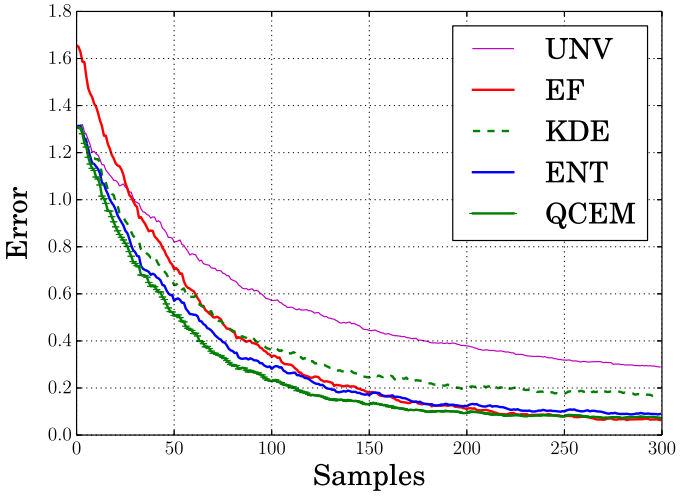


Fig. 4. Learning curves of each contrastivity approach.

provide the best sampling. Indeed, without any imposed moment constraints on the optimized distribution, the maximum entropy distribution is the uniform distribution. In a low dimensional setting the uniform distribution is an appropriate choice, but in high dimensions it quickly becomes problematic. Uniform sampling from a high dimensional volume is very inefficient, and a huge number of samples is required to ensure that we sample from regions where the data probability is sufficiently high. In contrast, QCEM combines the efficient sampling of data while providing homogeneous cover over the probability domain.

The full learning curves of all methods are depicted in Fig. 4. Consistent with the findings of [14], asymptotically, all approaches are able to find the underlying density. Nevertheless, Univariate and KDE convergence is much slower than the other methods and they are inappropriate sampling techniques for small datasets. The EF and ENT approaches have a moderate rate of convergence. Note that the ENT approach has slower convergence, presumably due to the numerical difficulties of entropy maximization [19]. QCEM objective avoids these numerical problems, yielding a faster alternative to these approaches.

3.2. Word embeddings similarity

Data and parameters. In the word embedding problem, the joint density over the sampled words and context pairs have to be learned. Following the state-of-the-art embedding evaluation schemas [25], we apply standard HTML text processing to Wikipedia. We remove words that occur less than 100 times in the whole corpus. This results in a sequence of several billion words, with a vocabulary size around 37k. The cooccurrence is then computed using windows of 10 tokens to each side of the focus word, following the practices of [2]. We use the word embedding architecture [16] that is known to be more robust for small sample sizes, dropouts and perturbations in the training set. The learning rate is initially set similarly to the methods and decayed in a linear fashion.

Evaluation and baselines. Despite the challenging nature of the objective evaluation of learned the word vectors, recent work in [25] suggests that intrinsic tasks, such as word similarity measurements, are a better proxy for measuring the generic quality of word vectors than the extrinsic evaluations. We therefore follow the experimental setup of [1,25], and compare the Spearman's correlation estimates of each model to human estimated similarities. Here a higher score indicates a higher correlation to human estimated word similarity judgements. For datasets containing multiple human annotators, we simply average the annotator scores. The WSS and WSR [1] are similarity and relatedness subsets of WordSim353 [10] dataset. WSS contains taxonomic relations (e.g. synonymy) and WSR mainly covers topical relations. These two datasets are relatively small and contain words that have relatively high frequency. MEN [3] word pair dataset contains 3k randomly sampled words, that occur at least 700 times, extracted from a freely available combined corpora having approximately 2.7B tokens. Sampling was performed to ensure balanced range of relatedness levels. The human similarity scores for this dataset are annotated using an interface for the Amazon Mechanical Turk. The RW [18] dataset contains 2034 word pairs, first word randomly sampled from Wikipedia documents. Then the outliers are filtered using WordNet entries, and the second word is sampled from synonym sets. Both MEN and RW contains many words with low frequency.

Baselines. We compare the following methods:

- *RG*, which uses the word frequency distribution, the data statistics, as its negative distribution p_n.
- *RGP*, uses a power heuristic of the unigram distribution. The powered version of the word frequencies are used $\sim p_c(\boldsymbol{w})^{\lambda}$. This heuristic is the common baseline that is used by the state-of-the-art method [20], where λ is a corpus dependent parameter. For a fair comparison, we set λ accordingly to the empirical findings of [20,23] as it is known to yield the best results for English corpora.
- *Uni* (Uniform) approach. We use a uniform distribution which all words of the vocabulary are equally probable to be picked as contrastive samples.
- *QCEM*, our proposed approach. For the problem construction, we use the unigram frequencies as data constraints: $\boldsymbol{d} \sim \Sigma_r \boldsymbol{C}_r$ which $\boldsymbol{C}_r$ are rows of word cooccurence matrix. For scalability considerations, we optimize over equivalence classes of words: defined such that words with the same frequency are in the same class. This equivalence strategy yields 5.2 k variables to optimize instead of 37k variables, increasing speed by an order of magnitude. Finally, we did not assume any *a priori* precision and decided to use an isotropic $\boldsymbol{\Sigma}$.

Quantitative results. Fig. 5 shows word similarity performances for all approaches on all datasets. In all datasets, the *RG* baseline performs poorly. For simpler datasets such as WSR and WSS, *QCEM* outperforms all baselines, especially on the lower dimensional regime where the correlation gain is slightly larger than high dimensional regime. Both taxonomy relations in WSS, and topical relations of WSR gain from the *QCEM* sampling. The *Uni* approach yields competitive performance especially in lower dimensions, but on high dimensional data the performance degrades quickly, as in the WSR and MEN datasets.

The performance gaps becomes more perceivable on more difficult datasets. On the MEN dataset, *RGP* is worse than *Uni* especially in lower dimensions, whereas in high dimensions the powering approach is better than the uniform distribution. *QCEM* and *Uni* perform quite alike in low dimensional MEN experiments. We believe this is due to two reasons. First, MEN similarity scores are much more noisy than other WS datasets due to the non-expert annotators which conceals the performance gap. Secondly, MEN vocabulary content is much broader than other datasets and it contains words occurring more than 700 times. This means query words are mostly from the heavy tail region in which *QCEM* and *Uni* behaves similarly. Nevertheless, *QCEM* does not suffer from performance losses in high dimensions like *Uni* approach and consistently achieves better performance.

On the RW dataset, it is noteworthy that the uniform contrast approach outperforms the powering heuristic with a small margin, for all model instances. For the WSS and WSR datasets, the powering heuristic obtains a reasonable performance whereas in

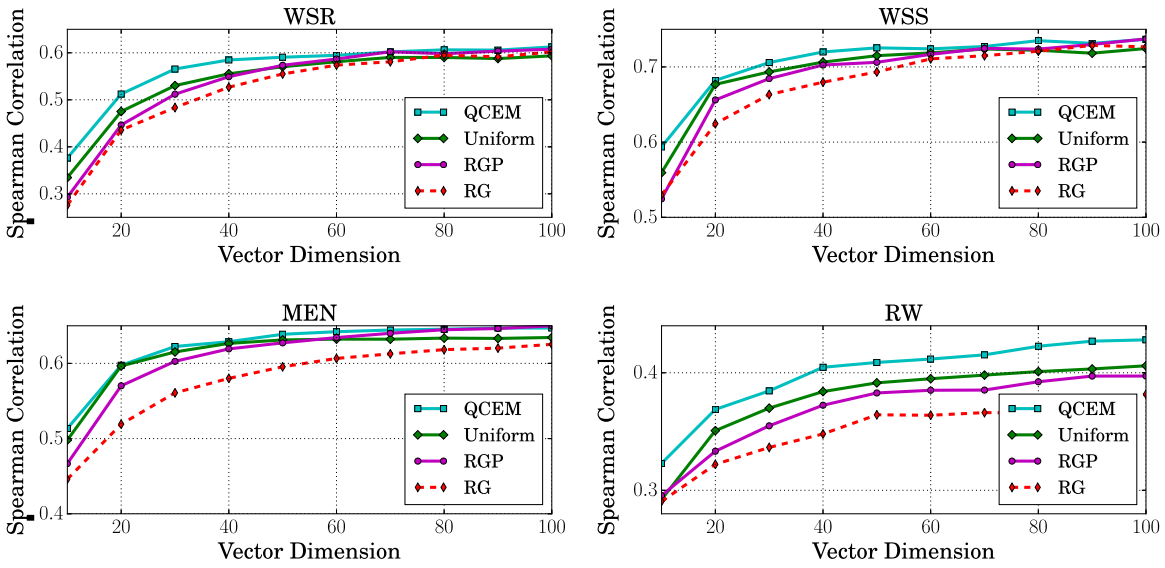


Fig. 5. Word Similarity performances of the methods on WSR, WSS, MEN and RW datasets.

the RW dataset it performs worse. Apparently, the constraints imposed by the powering heuristic turn out to be inappropriate for the RW dataset. This results in a suboptimal solution when the semantic relations of words are queried for a large set of less frequent words. The *QCEM* approach, on the other hand, does not impose such constraints, and obtains performance improvements with large margins. In the RW dataset, we finally compute the average correlation score over all the models, resulting in a 2.0% increase over the powering heuristic and a 4.8% over the standard baseline, a powerful quantitative indicator that embeddings trained with *QCEM* yield more realistic structure than the ones trained with computationally simple, but theoretically not justified heuristics.

3.3. Real world text classification

Setup. We evaluate vectors in the Agnews text classification benchmark, which consists of news articles collected from multiple sources. The dataset is randomly split into 120 k training and 7 k test documents and the goal is to predict the label of each document from {world,sports,business,science-technology} classes.

We plug in trained word vectors to a standard Multi Layer Perceptron (MLP) with logistic activation units and ensure fair comparison by fixing the embedding weights during the training which means the word vector layer does not change. This helps us to accurately quantify the performance gain from input vectors. Experiments are carried on with varying numbers of hidden units to evaluate how vectors contribute to different type of networks and whether they provide a sufficient generalization for different architectures. Each network is then trained using a standard Stochastic Gradient Descent optimizer with an adaptive learning schema. We then compute the F-1 scores for each approach.

Results. The result of each experiment is shown in Fig. 6. *RGP* approach performs worse in general. Networks trained with *RG* vectors occasionally perform well, but perform poorly on average. These vectors suffer from performance fluctuations suggesting

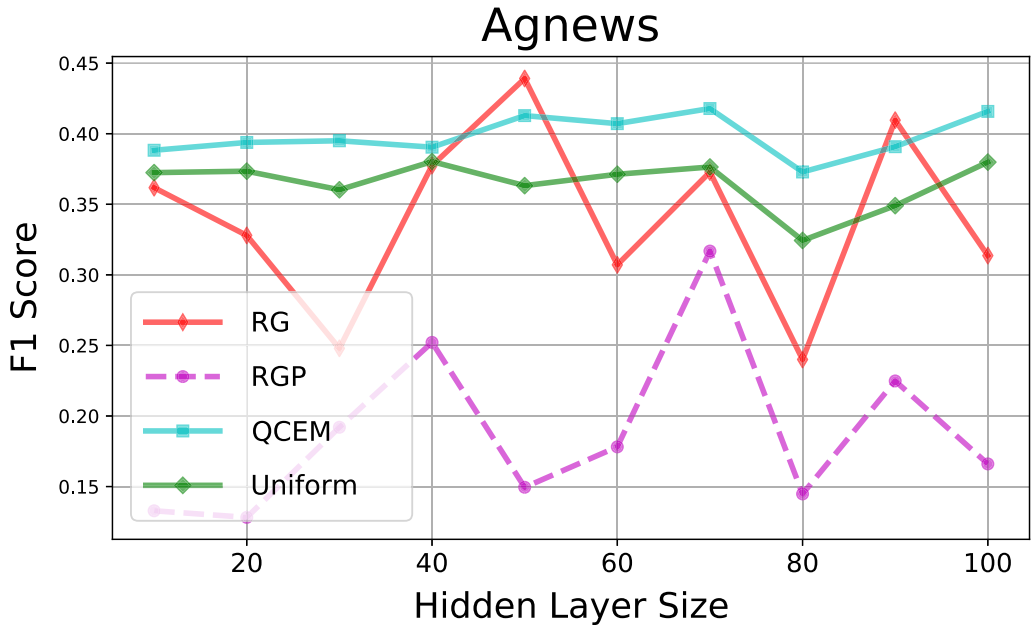
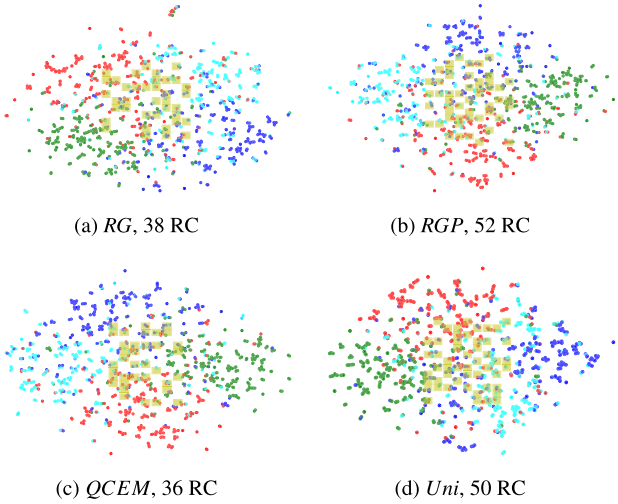


Fig. 6. Agnews text classification performance of vectors trained with each sampling algorithm.



(a) *RG*, 38 RC (b) *RGP*, 52 RC

(c) *QCEM*, 36 RC (d) *Uni*, 50 RC

Fig. 7. T-SNE dimensionality reduction of document embeddings. Each class is coded with a color. As confusions are common in the center region, we quantify the number of confusion regions (samples from multiple classes are present). Yellow boxes indicate Region of Confusions (RC). Less clutter is observed for *QCEM* embeddings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

that they are less robust to the local minima inherent in the problem. We also observe this phenomenon when we use vectors with *RGP*, illustrating another reason why optimizing the sampling distribution with our approach is advantageous. Note that performance of *QCEM* does not deteriorate even for networks with large number of neurons, and produces more stable scores. We visualize the documents constructed from embeddings. In Fig. 7, we show dimensionality-reduced document vectors in which each yellow region denotes a Region of Confusion. We expect document embeddings to have low intra-class distances, and high inter-class distances. *QCEM* document clusters are more coherent, and subject to less confusion in the center region.

4. Conclusions

We have presented a novel framework for optimizing negative sampling distribution using our Quadratically Constrained Entropy Maximization (QCEM) approach. Our formulation poses a convex and computationally tractable solution, has linear time complexity with respect to the vocabulary size, and permits scaling to large word embedding problems. Our theoretical analysis shows not only the generality but also the relation of our work to the prior heuristic state of the art approach, which is shown to be an approximation to our general maximization framework.

We validated our formulation both in synthetic density and real-world word vector space learning experiments, demonstrating that QCEM obtains faster convergence rates compared to a variety of competing approaches for learning exponential family probability densities. We reported the performance of QCEM in word similarity tasks, in which assumptions of the heuristic methods was not fulfilled. Results are shown for word similarity and text classification tasks, but implications of our framework extend to tasks such as Automatic Text Summarization [8,28]. In summary, QCEM can learn rare aspects of word meanings, especially when high sampling bias is present in the documents. As optimized distributions promote diversity, the chance to discover such aspects increases, especially when document size is limited. We leave validating our framework on ATS tasks to future work. Combination of the theoretical results and empirical evidence obtained for the vector space learning problems suggests that QCEM is an attractive solution to apply for determining negative sampling distributions.

Acknowledgments

This work was supported by the Dutch Organization for Scientific Research (NWO; grant 612.001.301).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2019.04.027

References

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and wordnet-based approaches, in: Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 19–27.
[2] S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, Random walks on context spaces: towards an explanation of the mysteries of semantic word embeddings, CoRR abs/1502.03520 (2015).
[3] M. Baroni, G. Dinu, G. Kruszewski, Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 238–247.
[4] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003).
[5] C.D. Boom, S.V. Canneyt, T. Demeester, B. Dhoedt, Representation learning for very short texts using weighted word embedding aggregation, Pattern Recognit. Lett. 80 (2016) 150–156.
[6] A. Bordes, S. Chopra, J. Weston, Question answering with subgraph embeddings, CoRR abs/1406.3676 (2014).
[7] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[8] D. Das, A.F.T. Martins, A survey on automatic text summarization, 2007,
[9] M. Fazel, T.K. Pong, D. Sun, P. Tseng, Hankel matrix rank minimization with applications to system identification and realization, SIAM J. Matrix Anal. Appl. 34 (2013) 946–977.
[10] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: The concept revisited, in: Proceedings of the 10th International Conference on World Wide Web, ACM, New York, NY, USA, 2001, pp. 406–414.
[11] J.R. Firth, A synopsis of linguistic theory 1930-55. 1952-59 (1957) 1–32.
[12] D. Ghosh, W. Guo, S. Muresan, Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words., in: Proceedings of the EMNLP, The Association for Computational Linguistics, 2015, pp. 1003–1012.
[13] M.U. Gutmann, A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, J. Mach. Learn. Res. 13 (2012) 307–361.
[14] M.U. Gutmann, A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, J. Mach. Learn. Res. 13 (2012) 307–361.
[15] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu, Exploring the limits of language modeling, CoRR abs/1602.02410 (2016).
[16] T. Kekec, D.M.J. Tax, Robust gram embeddings, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.
[17] O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization, in: Proceedings of the Advances in Neural Information Processing Systems 27, 2014, pp. 2177–2185.
[18] T. Luong, R. Socher, C.D. Manning, Better word representations with recursive neural networks for morphology, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL Sofia, Bulgaria, 2013, pp. 104–113. August 8–9
[19] R. Malouf, A comparison of algorithms for maximum entropy parameter estimation, in: Proceedings of the 6th Conference on Natural Language Learning, 20, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 1–7.
[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, CoRR abs/1310.4546 (2013).
[21] A. Mogadala, A. Rettinger, Bilingual word embeddings from parallel and non–parallel corpora for cross-language text classification., in: Proceedings of the HLT-NAACL, 2016, pp. 692–702.
[22] B. O'Donoghue, E. Chu, N. Parikh, S. Boyd, Conic optimization via operator splitting and homogeneous self-dual embedding, J. Optim. Theory Appl. 169 (2016) 1042–1068.
[23] J. Pennington, R. Socher, C. Manning, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.
[24] R.T. Rockafellar, Convex analysis, Princeton Mathematical Series, Princeton University Press, Princeton, N. J., 1970.
[25] T. Schnabel, I. Labutov, D.M. Mimno, T. Joachims, Evaluation methods for unsupervised word embeddings., in: Proceedings of the EMNLP, 2015, pp. 298–307.

[26] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.
[27] X. Tang, X. Wan, Learning bilingual embedding model for cross-language sentiment classification, in: Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2, 2014, pp. 134–141.
[28] J. Torres-Moreno, Automatic text summarization, Cognitive Science and Knowledge Management Series, Wiley, 2014.
[29] F. Vasile, E. Smirnova, A. Conneau, Meta-prod2vec: Product embeddings using side-information for recommendation, in: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2016, pp. 225–232.
[30] M.J. Wainwright, M.I. Jordan, et al., Graphical models, exponential families, and variational inference, Found. Trends® Mach. Learn. 1 (2008) 1–305.