Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization

Konstantin Vorontsov[1] and Anna Potapenko[2]

[1] Moscow Institute of Physics and Technology,
Dorodnicyn Computing Centre of RAS, The Higher School of Economics
voron@forecsys.ru
[2] Moscow State University, Dorodnicyn Computing Centre of RAS
anya_potapenko@mail.ru

Abstract. Probabilistic topic modeling of text collections is a powerful tool for statistical text analysis. In this tutorial we introduce a novel non-Bayesian approach, called *Additive Regularization of Topic Models*. ARTM is free of redundant probabilistic assumptions and provides a simple inference for many combined and multi-objective topic models.

Keywords: probabilistic topic modeling, regularization of ill-posed inverse problems, stochastic matrix factorization, Probabilistic Latent Sematic Analysis, Latent Dirichlet Allocation, EM-algorithm.

1 Introduction

Topic modeling is a rapidly developing branch of statistical text analysis [1]. Topic model uncovers a hidden thematic structure of the text collection and finds a highly compressed representation of each document by a set of its topics. From the statistical point of view, each topic is a set of words or phrases that frequently co-occur in many documents. The topical representation of a document captures the most important information about its semantics and therefore is useful for many applications including information retrieval, classification, categorization, summarization and segmentation of texts.

Hundreds of specialized topic models have been developed recently to meet various requirements coming from applications. For example, some of the models are capable to discover how topics evolve through time, how they are connected to each other, how they form topic hierarchies. Other models take into account additional information such as authors, sources, categories, citations or links between documents, or other kinds of document labels [2]. They can also be used to reveal the semantics of non-textual objects connected to the documents such as images, named entities or document users. Some of the models are focused on making topics more stable, sparse, robust, and better interpretable by humans. Linguistically motivated models benefit from syntactic considerations, grouping words into n-grams, finding collocations or constituent phrases. More ideas and applications of topic modeling can be found in the survey [3].

A *probabilistic topic model* defines each topic by a multinomial distribution over words, and then describes each document with a multinomial distribution over topics. Most recent models are based on a mainstream topic model LDA, Latent Dirichlet Allocation [4]. LDA is a two-level Bayesian generative model, which assumes that topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions. This assumption facilitates Bayesian inference due to the fact that the Dirichlet distribution is a conjugate to the multinomial one. However, the Dirichlet distribution has no convincing linguistic motivations and conflicts with two natural assumptions of sparsity: (1) most of the topics have zero probability in a document, and (2) most of the words have zero probability in a topic. The attempts to provide sparsity preserving Dirichlet prior lead to overcomplicated models [5,6,7,8,9]. Finally, Bayesian inference complicates the combination of many requirements into a single multi-objective topic model. The evolutionary algorithms recently proposed in [10] seem to be computationally infeasible for large text collections.

In this tutorial we present a survey of popular topic models in terms of a novel non-Bayesian approach — *Additive Regularization of Topic Models* (ARTM) [11], which removes the above limitations, simplifies theory without loss of generality, and reduces barriers to entry into topic modeling research field.

The motivations and essentials of ARTM may be briefly stated as follows. Learning of a topic model from a text collection is an ill-posed inverse problem of stochastic matrix factorization. Generally it has an infinite set of solutions. To choose a better solution we add a weighted sum of problem-oriented regularization penalty terms to the log-likelihood. Then the model inference in ARTM can be performed by a simple differentiation of the regularizers over model parameters. We show that many models, which previously required a complicated inference, can be obtained “in one line” within ARTM. The weights in a linear combination of regularizers can be adopted during the iterative process. Our experiments demonstrate that ARTM can combine regularizers that improve many criteria at once almost without a loss of the likelihood.

2 Topic models PLSA and LDA

In this section we describe Probabilistic Latent Sematic Analysis (PLSA) model, which was historically a predecessor of LDA. PLSA is a more convenient starting point for ARTM because it does not have regularizers at all. We provide the Expectation-Maximization (EM) algorithm with an elementary explanation, then describe an experiment on the model data that shows the instability of both PLSA and LDA models. The non-uniqueness and the instability of the solution does motivate a problem-oriented additive regularization.

Model assumptions. Let D denote a set (collection) of texts and W denote a set (vocabulary) of all words from these texts. Note that vocabulary may contain keyphrases as well, but we will not distinguish them from single words. Each document $d \in D$ is a sequence of n_d words $(w_1, \dots, w_{n_d})$ from the vocabulary W. Each word might appear multiple times in the same document.

Assume that each word occurrence in each document refers to some latent topic from a finite set of topics T. Text collection is considered to be a sample of triples $(w_i, d_i, t_i),\ i = 1, \dots, n$ drawn independently from a discrete distribution $p(w, d, t)$ over a finite probability space $W \times D \times T$. Words w and documents d are observable variables, while topics t are *latent* (hidden) variables.

Following the "bag of words" model, we represent each document by a subset of words $d \subset W$ and the corresponding integers n_{dw}, which count how many times the word w appears in the document d.

Conditional independence is an assumption that each topic generates words regardless of the document: $p(w \,|\, t) = p(w \,|\, d, t)$. According to the law of total probability and the assumption of conditional independence

$$p(w \,|\, d) = \sum_{t \in T} p(t \,|\, d)\, p(w \,|\, t). \tag{1}$$

The probabilistic model (1) describes how the collection D is generated from the known distributions $p(t \,|\, d)$ and $p(w \,|\, t)$. Learning a topic model is an inverse problem: to find distributions $p(t \,|\, d)$ and $p(w \,|\, t)$ given a collection D.

Stochastic matrix factorization. Our problem is equivalent to finding an approximate representation of observable data matrix

$$F = \bigl(f_{wd}\bigr)_{W \times D}, \qquad f_{wd} = \hat{p}(w \,|\, d) = n_{dw}/n_d,$$

as a product $F \approx \Phi\Theta$ of two unknown matrices — the matrix Φ of *word probabilities for the topics* and the matrix Θ of *topic probabilities for the documents*:

$$\begin{aligned} \Phi &= (\phi_{wt})_{W \times T}, & \phi_{wt} &= p(w \,|\, t), & \phi_t &= (\phi_{wt})_{w \in W}; \\ \Theta &= (\theta_{td})_{T \times D}, & \theta_{td} &= p(t \,|\, d), & \theta_d &= (\theta_{td})_{t \in T}. \end{aligned}$$

Matrices F, Φ and Θ are *stochastic*, that is, their columns f_d, ϕ_t, θ_d are non-negative and normalized representing discrete distributions. Usually the number of topics $|T|$ is much smaller than both $|D|$ and $|W|$.

Likelihood maximization. In probabilistic latent semantic analysis (PLSA) [12] the topic model (1) is learned by the log-likelihood maximization:

$$\ln \prod_{i=1}^{n} p(d_i, w_i) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \,|\, d) + \sum_{d \in D} n_d \ln p(d) \to \max,$$

which results in a constrained maximization problem:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \ \to\ \max_{\Phi, \Theta}; \tag{2}$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \tag{3}$$

Algorithm 2.1: The rational EM-algorithm for PLSA.

```
Input: document collection $D$, number of topics $|T|$, initialized $\Phi$, $\Theta$;
Output: $\Phi$, $\Theta$;
repeat
    zeroize $n_{wt}$, $n_{dt}$, $n_t$, $n_d$ for all $d \in D,\ w \in W,\ t \in T$;
    for all $d \in D,\ w \in d$
        $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
        for all $t \in T$: $\phi_{wt}\theta_{td} > 0$
            increase $n_{wt}$, $n_{dt}$, $n_t$, $n_d$ by $\delta = n_{dw}\phi_{wt}\theta_{td}/Z$;
    $\phi_{wt} := n_{wt}/n_t$ for all $w \in W,\ t \in T$;
    $\theta_{td} := n_{dt}/n_d$ for all $d \in D,\ t \in T$;
until $\Phi$ and $\Theta$ converge;
```

EM-algorithm. The problem (2), (3) can be solved by an iterative EM-algorithm. First, the columns of the matrices Φ and Θ are initialized with random distributions. Then two steps (E-step and M-step) are repeated in a loop.

At the E-step the probability distributions for the latent topics $p(t \mid d, w)$ are estimated for each word w in each document d using the Bayes' rule. Auxiliary variables n_{dwt} are introduced to estimate how many times the word w appears in the document d with relation to the topic t:

$$n_{dwt} = n_{dw} p(t \mid d, w), \quad p(t \mid d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}. \tag{4}$$

At the M-step summation of n_{dwt} values over d, w, t provides empirical estimates for the unknown conditional probabilities:

$$\begin{aligned} \phi_{wt} &= \frac{n_{wt}}{n_t}, & n_{wt} &= \sum_{d \in D} n_{dwt}, & n_t &= \sum_{w \in W} n_{wt}, \\ \theta_{td} &= \frac{n_{dt}}{n_d}, & n_{dt} &= \sum_{w \in d} n_{dwt}, & n_d &= \sum_{t \in T} n_{dt}, \end{aligned}$$

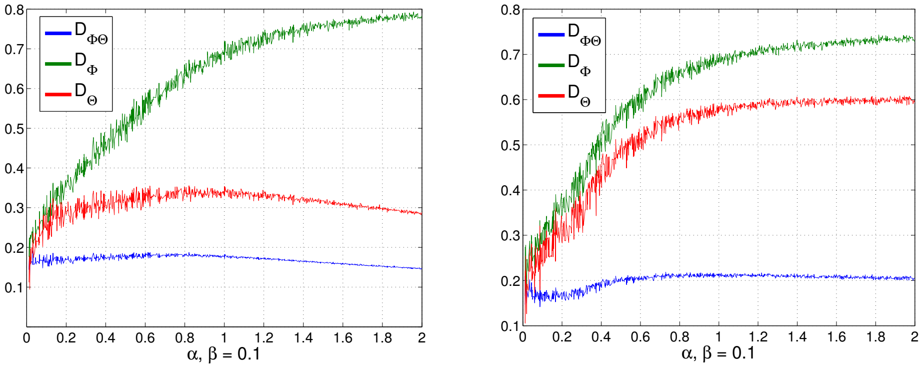
which can be rewritten in a shorter notation using the proportionality sign $\propto$:

$$\phi_{wt} \propto n_{wt}, \qquad \theta_{td} \propto n_{dt}. \tag{5}$$

Equations (4), (5) define a necessary condition for a local optimum of the problem (2), (3). In the next section we will prove this for a more general case.

The system of equations (4), (5) can be solved by various numerical methods. The simple iteration method leads to a family of *EM-like* algorithms, which may differ in implementation details. For example, Algorithm 2.1 avoids storing the three-dimensional array n_{dwt} by incorporating the E-step inside the M-step.

Latent Dirichlet Allocation. In LDA parameters Φ, Θ are constrained to avoid overfitting [4]. LDA assumes that the columns of the matrices Φ and Θ are



LDA PLSA

Fig. 1. Errors in restoring the matrices Φ, Θ and $\Phi\Theta$ over hyperparameter α ($\beta = 0.1$).

drawn from the Dirichlet distributions with positive vectors of hyperparameters $\beta = (\beta_w)_{w \in W}$ and $\alpha = (\alpha_t)_{t \in T}$ respectively.

Learning algorithms for LDA generally fall into two categories — sampling-based algorithms [13] or variational algorithms [14]. They can be considered also as EM-like algorithms with modified M-step [15]. The following is the most simple and frequently used modification:

$$\phi_{wt} \propto n_{wt} + \beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_t. \tag{6}$$

This modification has the effect of smoothing, since it increases small probabilities and decreases large probabilities.

The non-uniqueness problem. The likelihood (2) depends on the product $\Phi\Theta$, not on separate matrices Φ and Θ. Therefore, for any linear transformation S such that matrices $\Phi' = \Phi S$ and $\Theta' = S^{-1}\Theta$ are stochastic, their product $\Phi'\Theta' = \Phi\Theta$ gives the same value of the likelihood. The transformation S depends on a random initialization of the EM-algorithm. Thus, learning a topic model is an ill-posed problem whose solution is not unique and hence is not stable.

The following experiment on the model data verifies the ability of PLSA and LDA to restore true matrixes Φ, Θ. The collection was generated with the size parameters $|W| = 1000, \;\; |D| = 500, \;\; |T| = 30$. The lengths of the documents $n_d \in [100, 600]$ were chosen randomly. Columns of the matrices Φ, Θ were drawn from the symmetric Dirichlet distributions with parameters β, α respectively. The differences between the restored distributions $\hat{p}(i \,|\, j)$ and the model ones $p(i \,|\, j)$ were measured by the average Hellinger distance both for the matrices Φ, Θ and for their product:

$$D_\Phi = H(\hat{\Phi}, \Phi); \quad D_\Theta = H(\hat{\Theta}, \Theta); \quad D_{\Phi\Theta} = H(\hat{\Phi}\hat{\Theta}, \Phi\Theta);$$
$$H(\hat{p}, p) = \frac{1}{m} \sum_{j=1}^{m} \biggl(\frac{1}{2} \sum_{i=1}^{n} \Bigl(\sqrt{\hat{p}(i \,|\, j)} - \sqrt{p(i \,|\, j)} \Bigr)^2 \biggr)^{\frac{1}{2}}.$$

Both PLSA and LDA restore $\varPhi$ and $\varTheta$ much worse than their product, Fig. 1. The error are less for sparse original matrices $\varPhi, \varTheta$. LDA did not perform well even when the same α, β are used for both generating and restoring stages.

This experiment shows that the Dirichlet regularization can not ensure a stable solution. Stronger regularizer or combination of regularizers should be used.

Also we conclude that PLSA model being free of any regularizers is the most convenient starting point for multi-objective problem-oriented regularization.

3 Additive regularization for topic models

In this section we introduce the additive regularization framework and prove a general equation for a regularized M-step in the EM-algorithm.

Consider r objectives $R_i(\varPhi, \varTheta),\ \ i = 1, \dots, r$, called *regularizers*, which have to be maximized together with the likelihood (2). According to a standard scalarization approach to the multi-objective optimization we maximize a linear combination of the objectives L and R_i with nonnegative *regularization coefficients* τ_i:

$$R(\varPhi, \varTheta) = \sum_{i=1}^{r} \tau_i R_i(\varPhi, \varTheta), \qquad L(\varPhi, \varTheta) + R(\varPhi, \varTheta) \ \to \ \max_{\varPhi, \varTheta}. \tag{7}$$

Topic t is called *overregularized* if $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leq 0$ for all words $w \in W$.

Document d is called *overregularized* if $n_{dt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leq 0$ for all topics $t \in T$.

Theorem 1. *If the function $R(\varPhi, \varTheta)$ is continuously differentiable and $(\varPhi, \varTheta)$ is the local minimum of the problem* (7), (3)*, then for any topic t and any document d that are not overregularized the system of equations holds:*

$$n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \tag{8}$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_{+}; \qquad n_{wt} = \sum_{d \in D} n_{dwt}; \tag{9}$$

$$\theta_{td} \propto \left(n_{dt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_{+}; \qquad n_{dt} = \sum_{w \in d} n_{dwt}; \tag{10}$$

where $(z)_+ = \max\{z, 0\}$.

Note 1. Equation (9) gives $\phi_t = 0$ for overregularized topics t. Equation (10) gives $\theta_d = 0$ for overregularized documents d. Overregularization is an important mechanism, which helps to exclude insignificant topics and documents out of the topic model. Regularizers that encourage topic exclusions may be used to optimize the number of topics. A document may be excluded if it is too short or does not contain topical words.

Note 2. The system of equations (8)–(10) defines a regularized EM-algorithm. It keeps E-step from (4) and redefines M-step by regularized equations (9), (10). If $R(\varPhi, \varTheta) = 0$ then the regularized topic model is reduced to the usual PLSA.

Proof. For the local minimum $(\varPhi, \varTheta)$ of the problem (7), (3) the KKT conditions (see Appendix A) can be written as follows:

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w \,|\, d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geq 0; \quad \lambda_{wt}\phi_{wt} = 0.$$

Let us multiply both sides of the first equation by ϕ_{wt}, reveal the auxiliary variable n_{dwt} from (8) in the left-hand side and sum it over d:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w \,|\, d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

An assumption that $\lambda_t \leq 0$ contradicts the condition that topic t is not overregularized. Then $\lambda_t > 0$, $\phi_{wt} \geq 0$, the left-hand side is nonnegative, thus the right-hand side is nonnegative too, consequently,

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right)_+. \tag{11}$$

Let us sum both sides of this equation over all $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right)_+. \tag{12}$$

Finally, we obtain (9) by expressing ϕ_{wt} from (11) and (12).

Equations for θ_{td} can be derived analogously thus finalizing the proof.

The EM-algorithm for learning regularized topic models can be implemented by easy modification of any EM-like algorithm at hand. In Algorithm 2.1 only steps 7 and 8 are to be modified according to equations (9) and (10).

4 A survey of regularizers for topic models

In this section we revisit some of the well known topic models and show that ARTM significantly simplifies their inference and modifications. We propose an alternative interpretation of LDA as a regularizer that minimizes KL-divergence with a fixed distribution. Then we revisit topic models for sparsing domain-specific topics, smoothing background (common lexis) topics, semi-supervised learning, number of topics optimization, topics decorrelation, topic coherence maximization, documents linking, and document classification. We also consider the problem of combining regularizers and introduce the notion of *regularization trajectory*.

Smoothing regularization and LDA. Let us minimize the KL-divergence (see Appendix B) between the distributions ϕ_t and a fixed distribution $\beta = (\beta_w)_{w\in W}$, and the KL-divergence between θ_d and a fixed distribution $\alpha = (\alpha_t)_{t\in T}$:

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}, \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

After summing these criteria with coefficients β_0, α_0 and removing constants we have the regularizer

$$R(\Phi, \Theta) = \beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to \max.$$

The regularized M-step (9) and (10) gives us two equations

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \qquad \theta_{td} \propto n_{dt} + \alpha_0 \alpha_t,$$

which are exactly the same as the M-step (6) in LDA model with hyperparameter vectors $\beta = \beta_0(\beta_w)_{w\in W}$ and $\alpha = \alpha_0(\alpha_t)_{t\in T}$ of the Dirichlet distributions.

The non-Bayesian interpretation of the smoothing regularization in terms of KL-divergence is simple and natural. Moreover, it avoids complicated inference techniques such as Variational Bayes or Gibbs Sampling.

Sparsing regularization. The opposite regularization strategy is to maximize KL-divergence between ϕ_t, θ_d and fixed distributions β, α:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to \max.$$

For example, to find a sparse distributions ϕ_{wt} with lower entropy we may choose the uniform distribution $\beta_w = \frac{1}{|W|}$, which is known to have the largest entropy.

The regularized M-step (9) and (10) gives equations that differ from the smoothing equations only in the sign of the parameters β, α:

$$\phi_{wt} \propto \bigl(n_{wt} - \beta_0 \beta_w\bigr)_+, \qquad \theta_{td} \propto \bigl(n_{dt} - \alpha_0 \alpha_t\bigr)_+.$$

The idea of entropy-based sparsing was originally proposed in the dynamic PLSA for video processing tasks [16] to produce sparse distributions of topics over time. The Dirichlet prior conflicts with sparsing assumption, which leads to sophisticated sparse LDA models [5,6,7,8,9]. Simple and natural sparsing is possible only by abandoning the Dirichlet prior assumption.

Combining smoothing and sparsing. In modeling a multidisciplinary text collection topics should contain domain-specific words and be free of common lexis words. To learn such a model we suggest to split the set of topics T into two subsets: sparse domain-specific topics S and smoothed background topics B. Background topics should be close to a fixed distribution over words β_w and should appear in all documents. The model with background topics B is an extension of robust models [17,18], which used a single background distribution.

Semi-supervised learning. Additional training data can further improve quality and interpretability of a topic model. Assume that we have a prior knowledge, stating that each document d from a subset $D_0 \subseteq D$ is associated with a subset of topics $T_d \subset T$. Analogically, assume that each topic $t \in T_0$ contains a subset of words $W_t \subset W$. Consider a regularizer that maximizes the total probability of topics in T_d and the total probability of words in W_t:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta_{td} \to \max.$$

The regularized M-step (9) and (10) gives yet another sort of smoothing:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi_{wt},\ t \in T_0,\ w \in W_t; \quad \theta_{td} \propto n_{dt} + \alpha_0 \theta_{td},\ d \in D_0,\ t \in T_d.$$

Sparsing regularization of topic probabilities for the words $p(t \,|\, d, w)$ is motivated by a natural assumption that each word in a text is usually related to one topic. To meet this requirement we use the entropy-based sparsing and maximize the average KL-divergence between $p(t \,|\, d, w)$ and uniform distribution over topics:

$$\sum_{d,w} n_{dw} \, \mathrm{KL}\bigl(\tfrac{1}{|T|} \,\|\, p(t \,|\, d, w)\bigr) \to \min_{\Phi,\Theta};$$

$$R(\Phi, \Theta) = \frac{\tau}{|T|} \sum_{d,w} n_{dw} \sum_{t \in T} \ln \frac{\sum_{s \in T} \phi_{ws} \theta_{sd}}{\phi_{wt} \theta_{td}} \to \max.$$

The regularized M-step (9) and (10) gives

$$\phi_{wt} \propto \bigl(n_{wt} + \tau\bigl(n_{wt} - \tfrac{1}{|T|} n_w\bigr)\bigr)_+, \qquad \theta_{td} \propto \bigl(n_{dt} + \tau\bigl(n_{dt} - \tfrac{1}{|T|} n_d\bigr)\bigr)_+.$$

These equations mean that ϕ_{wt} decreases (and may eventually turn to zero) if the word w occurs in the topic t less frequently than in the average over all topics. Analogously, θ_{td} decreases (and may also turn to zero) if the topic t occurs in the document d less frequently than in the average over all topics.

Elimination of insignificant topics can be done by entropy-based sparsing of the global distribution over topics $p(t) = \sum_d p(d) \theta_{td}$. To do this we maximize the KL-divergence between $p(t)$ and the uniform distribution over topics:

$$R(\Theta) = \tau \sum_{t \in T} \ln \sum_{d \in D} p(d) \theta_{td} \to \max.$$

The regularized M-step (10) gives

$$\theta_{td} \propto \Bigl(n_{dt} - \tau \frac{n_d}{n_t} \theta_{td}\Bigr)_+.$$

This regularizer works as a row sparser for the matrix Θ because of n_t counter in the denominator. If n_t is small then the big values are subtracted from all elements n_{dt} of the t-th row of the matrix Θ. If all elements of a row will be set to zero then the corresponding topic t could never be used, i.e. it will be eliminated from the model. We can decrease the current number of active topics gradually during EM-iterations by increasing a coefficient τ until some of the quality measures will not deteriorate.

Note that this approach to the number of topics optimization is much simpler than the state-of-the-art Bayesian techniques such as Hierarchical Dirichlet Process [19] and Chinese Restaurant Process [20].

Covariance regularization for topics. Reducing the overlapping between the topic-word distributions is known to make the learned topics more interpretable [21]. A regularizer that minimizes covariance between vectors ϕ_t,

$$R(\Phi) = -\tau \sum_{t\in T} \sum_{s\in T\backslash t} \sum_{w\in W} \phi_{wt}\phi_{ws} \to \max,$$

leads to the following equation of the M-step:

$$\phi_{wt} \propto \Bigl(n_{wt} - \tau\phi_{wt} \sum_{s\in T\backslash t} \phi_{ws}\Bigr)_+.$$

That is, for each word w the highest probabilities ϕ_{wt} will increase from iteration to iteration, while small probabilities will decrease, and may eventually turn into zeros. Therefore, this regularizer also stimulates sparsity. Besides, it has another useful property, which is to group stop-words into separate topics [21].

Covariance regularization for documents. Sometimes we possess an information that some documents are likely to share similar topics. For example, they may fall into the same category or one document may have a reference or a link to the other. Making use of this information in terms of the regularizer, we get:

$$R(\Theta) = \tau \sum_{d,c} n_{dc} \sum_{t\in T} \theta_{td}\theta_{tc} \to \max,$$

where n_{dc} is the weight of the link between documents d and c. A similar LDA-JS model is described in [22], which is based on the minimization of Jensen–Shannon divergence between θ_d and θ_c, rather than on the covariance maximization.

According to (10), the equation for θ_{td} in the M-step turns into

$$\theta_{td} \propto n_{dt} + \tau\theta_{td} \sum_{c\in D} n_{dc}\theta_{tc}.$$

Thus the iterative process adjusts probabilities θ_{td} so that they become closer to θ_{tc} for all documents c, connected with d.

Coherence maximization. A topic is called *coherent* if the most frequent words from this topic typically appear nearby in the documents (either in the training collection, or in some external corpus like Wikipedia). An average topic coherence is known to be a good measure of interpretability of a topic model [23].

Consider a regularizer, which augments probabilities of coherent words [24]:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max,$$

where $C_{uv} = N_{uv} \bigl[\mathrm{PMI}(u,v) > 0\bigr]$ is the co-occurrence estimate of word pairs $(u,v) \in W^2$, pointwise mutual information $\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ is defined through document frequencies: N_{uv} is the number of documents that contain both words u, v in a sliding window of ten words, N_u is the number of documents that contain at least one occurrence of the word u.

Note that there is no common approach to the coherence optimization in the literature. Another coherence optimizer was proposed in [25] for LDA model and Gibbs Sampling algorithm with more complicated motivations through a generalized Polya urn model and a more complex heuristic estimate for C_{wv}. Again, this regularizer can be much easier reformulated in terms of ARTM.

The classification regularizer. Let C be a finite set of classes. Suppose each document d is labeled by a subset of classes $C_d \subset C$. The task is to infer a relationship between classes and topics, improve a topic model by using labels information, and to learn a decision rule to classify new documents. Common discriminative approaches such as SVM or Logistic Regression usually give unsatisfactory results on large text collections with a big number of unbalanced and interdependent classes. Probabilistic topic models can benefit in this situation [2].

Recent research papers provide various examples of document labeling. Classes may refer to text categories [2,26], authors [27], time periods [28,16], cited documents [22], cited authors [29], users of documents [30]. Many specialized models has been developed for these and other cases, more information can be found in surveys [3,2]. All these models fall into a small number of types that can be easily expressed in terms of ARTM. Below we consider one of the most general topic model for document classification.

Let us expand the probability space to the set $D \times W \times T \times C$ and assume that each word w in each document d is not only related to a topic $t \in T$, but also to a class $c \in C$. To classify documents we model a distribution $p(c \,|\, d)$ over classes for each document d. As in the Dependency LDA topic model [2], we assume that $p(c \,|\, d)$ is expressed in terms of distributions $p(c \,|\, t) = \psi_{ct}$ and $p(t \,|\, d) = \theta_{td}$ in a way, similar to the basic topic model (1):

$$p(c \,|\, d) = \sum_{t \in T} \psi_{ct} \theta_{td},$$

where $\Psi = (\psi_{ct})_{C \times T}$ is a new model parameters matrix. Our regularizer minimize KL-divergence between the probability model of classification $p(c \,|\, d)$ and the empirical frequency $m_{dc} = n_d \frac{[c \in C_d]}{|C_d|}$ of classes in the documents:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max.$$

The problem is still solved via EM-like algorithms. In addition to (4), the E-step estimates conditional probabilities $p(t \,|\, d, c)$ and auxiliary variables m_{dct}:

$$m_{dct} = m_{dc} p(t \,|\, d, c), \qquad p(t \,|\, d, c) = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T} \psi_{cs}\theta_{sd}}.$$

In the M-step ϕ_{wt} are estimated from (5), the estimates for ψ_{ct} are analogous to ϕ_{wt}, the estimates for θ_{td} accumulate counters of words and classes within the documents:

$$\psi_{ct} \propto m_{ct}, \;\; m_{ct} = \sum_{d\in D} m_{dct}; \qquad \theta_{td} \propto n_{dt} + \tau m_{dt}, \;\; m_{dt} = \sum_{c\in C} m_{dct}.$$

Additional regularizers for Ψ can be used to control sparsity.

Label regularization improves classification for multi-label classification problems with unbalanced classes [2] by minimizing KL-divergence between the model distribution $p(c)$ over classes and the empirical frequencies of classes $\hat{p}_c$ observed in the training data:

$$R(\Psi) = \tau \sum_{c\in C} \hat{p}_c \ln p(c) \to \max; \qquad p(c) = \sum_{t\in T} \psi_{ct} p(t), \quad p(t) = \frac{n_t}{n}.$$

The formula for the M-step is therefore as follows:

$$\psi_{ct} \propto m_{ct} + \tau \hat{p}_c \frac{\psi_{ct} n_t}{\sum_{s\in T} \psi_{cs} n_s}.$$

Regularization trajectory. A linear combination of multiple regularizers R_i depends on regularization coefficients τ_i, which require a special handling in practice. A similar problem is efficiently solved in ElasticNet algorithm, which combines L_1 and L_2-regularizers for regression and classification tasks [31]. In topic modeling there are far more various regularizers and they can influence each other in a non-trivial way. Our experiments show that some regularizers may worsen the convergence if they are activated too early or too abruptly. Therefore our recommendation is to choose the regularization trajectory experimentally.

5 Quality measures for topic models

The accuracy of a topic model $p(w \,|\, d)$ on the collection D is commonly evaluated in terms of *perplexity* closely related to the likelihood

$$\mathscr{P}(D, p) = \exp\Bigl(-\frac{1}{n} L(\Phi, \Theta)\Bigr) = \exp\biggl(-\frac{1}{n} \sum_{d\in D} \sum_{w\in d} n_{dw} \ln p(w \,|\, d)\biggr).$$

The *hold-out perplexity* $\mathscr{P}(D', p_D)$ of the model p_D trained on the collection D is evaluated on the test set of documents D', which does not overlap with D. In our experiments we split the collection randomly so that

$|D| : |D'| = 10 : 1$. Each testing document d is further randomly split into two halves: the first one is used to estimate parameters θ_d, and the second one is used in the perplexity evaluation. The words in the second halves that did not appear in D are ignored. Parameters ϕ_t are estimated from the training set.

The *sparsity* of a model is measured by the percent of zero elements in matrices Φ and Θ. For the models that separate domain-specific topics S and background topics B we estimate sparsity over domain-specific topics S only.

The high *ratio of background words* over document collection

$$\mathrm{BackgroundRatio} = \frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t \,|\, d, w)$$

may indicate the model degradation as a result of excessive sparsing or topics elimination and can be used as a stopping criterion for sparsing.

The *interpretability* of a topic model is evaluated indirectly by coherence, which is known to correlate well with human interpretability [32,23,25]. The *coherence of a topic* is defined as the *pointwise mutual information* averaged over all pairs of words within the k most probable words of the topic t:

$$\mathrm{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j)$$

where w_i is the i-th word in the list of ϕ_{wt}, $w \in W$, sorted in descending order. *Coherence of a topic model* is defined as average PMI_t over all domain-specific topics $t \in S$. In most papers the value k is fixed to 10. Due to a particular importance of the topic coherence we have also examined two additional measures: the coherence for $k = 100$, and the coherence for the topic kernels.

We define the *kernel* of each topic as a set of words that distinguish this topic from other topics: $W_t = \{w \colon p(t \,|\, w) > \delta\}$. In our experiments we set $\delta = 0.25$. We suggest that well interpretable topic must have a reasonable *kernel size* $|W_t|$ about 20–200 words and a high values of topic *purity* and *contrast*:

$$\mathrm{Purity}_t = \sum_{w \in W_t} p(w \,|\, t); \qquad \mathrm{Contrast}_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \,|\, w).$$

We define the corresponding measures of the overall topic model (kernel size, purity and contrast) by averaging over all domain-specific topics $t \in S$.

6 Experiments with combining regularizers

We are going to demonstrate ARTM approach in practice by combining regularizers for sparsing, smoothing, topics decorrelation, and number of topics optimization. Our objective is to build a highly sparse topic model with a better interpretability of topics, and at the same time to extract stop-words and common lexis words. Thus, we aim to improve several quality measures with no significant loss of the likelihood or perplexity.

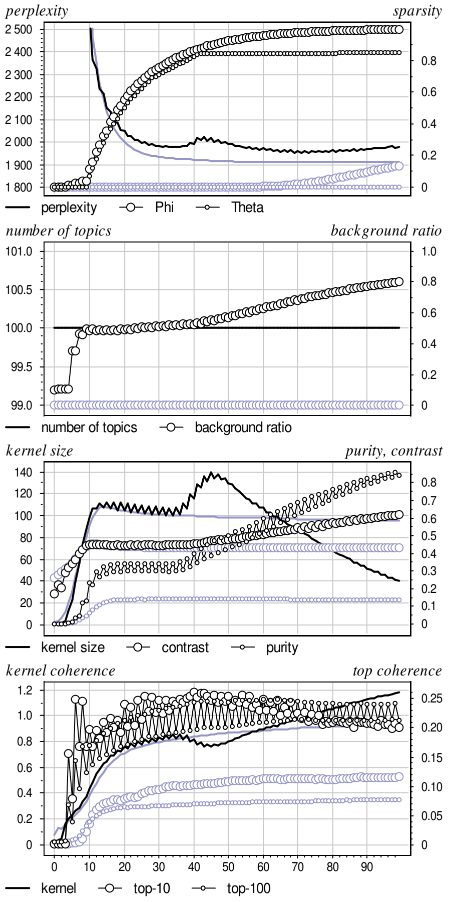


Fig. 2. Comparing PLSA (grey) vs. ARTM with sparsing, smoothing, and decorrelation (black).

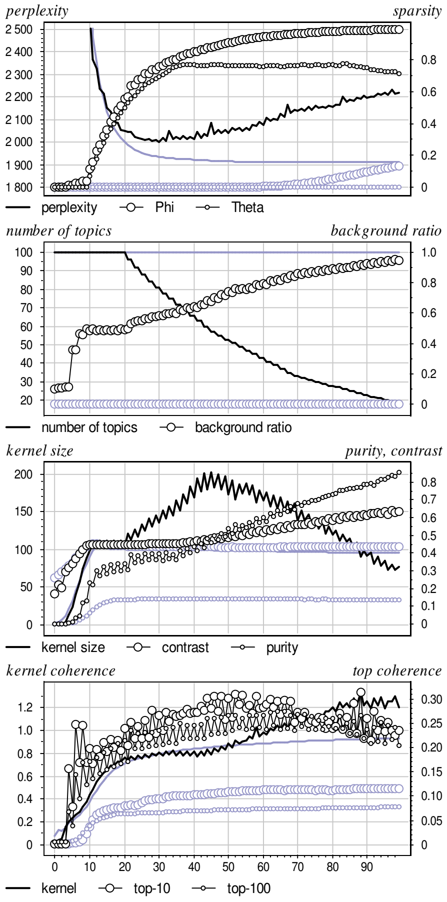


Fig. 3. Comparing PLSA (grey) vs. ARTM with sparsing, smoothing, decorrelation, and topics elimination (black).

Text collection. In our experiments we use the NIPS dataset, which contains $|D| = 1566$ English articles from the Neural Information Processing Systems conference. The length of the collection in words is $n \approx 2.3 \cdot 10^6$. The vocabulary size is $|W| \approx 1.3 \cdot 10^4$. The testing set has $|D'| = 174$ documents.

In the preparation step we used BOW toolkit [33] to perform changing to low-case, punctuation elimination, and stop-words removal.

In all the experiments the number of iterations was set to 100, and the number of topics was set to $|T| = 100$ with $|B| = 10$ background topics.

Experimental results. Figures 2–3 present quality measures of the topic model as a function of the iteration step. In each figure we compare two models, PLSA being shown with grey lines and ARTM with black lines.

Quality measures are shown in four charts, stack on top of each other in one column with synchronized horizontal axes. Top chart: perplexity on the left-hand axis, and sparsity of matrices Φ, Θ on the right-hand axis. Second chart: number of topics on the left-hand axis, and ratio of background words on the right-hand axis. Third chart: kernel size on the left-hand axis, and contrast and purity on the right-hand axis. Bottom chart: kernel coherence on the left-hand axis, and top10 and top100 coherence on the right-hand axis.

ARTM allows to use regularizers in any combination. Therefore, we explore how various combinations of regularizer influence different quality measures.

PLSA and LDA have performed similarly by all measures: perplexity is about 1900; sparsity is 0%; kernel size is 80–100 words; purity is 12%; contrast is 43%; coherence top10: 0.07, top100: 0.12, kernel: 0.9.

In ARTM we augment the regularization coefficient for sparsing gradually from the 10-th iteration. An earlier or a more abrupt sparsing may lead to perplexity deterioration. The gradual sparsing results in a highly sparse Φ matrix (98% of zeros) and Θ matrix (85% of zeros), while the perplexity becomes slightly worse. We smooth the background topics from the first iteration using the uniform distribution $\beta_w = 1/|W|$ and parameters $\alpha = 0.8,\ \ \beta = 0.1$. Using a non-uniform distribution $\beta_w = n_w/n$ yields similar results.

The decorrelation regularizer works well if activated from the very beginning. It does not change the perplexity significantly, and improves purity and coherence. Contrast and kernel size remain the same. However, the sparsity of Φ stays at 40%, which apparently is not good enough, and Θ does not get sparse at all. The combination of sparsing, smoothing and decorrelation provides the best results, shown in Fig. 2. Notice that in all experiments kernel coherence is considerably higher than top10 and top100 coherence.

The sparsing regularizer for insignificant topics elimination turned out to be in conflict with decorrelation. Therefore we apply decorrelation at even iterations, and topics elimination at odd iterations. In our experiments the removal of topics begins to deteriorate the model perplexity when the number of topics becomes less than 60, Fig. 3.

7 Conclusions

This tutorial gives a brief survey of topic models from a new non-Bayesian viewpoint which we call ARTM — *Additive Regularization of Topic Models.* ARTM makes topic models easy to design, easy to infer, and easy to explain. Many topic models are based on stochastic matrix factorization — an ill-posed inverse problem whose solution is non-unique and instable. The goal of regularization is to reduce a potentially infinite set of solutions, and to select a better one, which satisfies our additional requirements. These requirements can be formalized through a maximization of a weighted sum of regularizers, differentiable with respect to the parameters of the model. The EM-algorithm with a modified M-step can be used to solve the optimization problem. Our interpretation of the EM-algorithm is also nonprobabilistic. We consider the EM-algorithm as a simple iteration method for solving a system of equations that defines a necessary conditions of the local optimum. Problems of a numerical convergence and regularization trajectories are left beyond the scope of this paper.

Acknowledgements. The work was supported by the Russian Foundation for Basic Research grants 14-07-00847, 14-07-00908. We thank Alexander Frey for his help and valuable discussion, and Vitaly Glushachenkov for his experimental work on model data.

References

1. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4) (2012) 77–84
2. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Machine Learning **88**(1-2) (2012) 157–208
3. Daud, A., Li, J., Zhou, L., Muhammad, F.: Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China **4**(2) (2010) 280–301
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
5. Shashanka, M., Raj, B., Smaragdis, P.: Sparse overcomplete latent variable decomposition of counts data. In Platt, J.C., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems, NIPS-2007. MIT Press, Cambridge, MA (2008) 1313–1320
6. Wang, C., Blei, D.M.: Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In: NIPS, Curran Associates, Inc. (2009) 1982–1989
7. Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In: ICML'11. (2011) 1041–1048
8. Larsson, M.O., Ugander, J.: A concave regularization technique for sparse mixture models. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K., eds.: Advances in Neural Information Processing Systems 24. (2011) 1890–1898
9. Chien, J.T., Chang, Y.L.: Bayesian sparse topic model. Journal of Signal Processessing Systems (2013) 1–15
10. Khalifa, O., Corne, D., Chantler, M., Halley, F.: Multi-objective topic modelling. In: 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013), Springer LNCS (2013) 51–65
11. Vorontsov, K.V.: Additive regularization for topic models of text collections. Doklady Mathematics **88**(3) (2014) (to appear)
12. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1999) 50–57
13. Wang, Y.: Distributed Gibbs sampling of latent dirichlet allocation: The gritty details (2008)
14. Teh, Y.W., Newman, D., Welling, M.: A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In: NIPS. (2006) 1353–1360
15. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence. (2009) 27–34

16. Varadarajan, J., Emonet, R., Odobez, J.M.: A sparsity constraint for topic models — application to temporal activity mining. In: NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. (2010)
17. Chemudugunta, C., Smyth, P., Steyvers, M. In: Modeling general and specific aspects of documents with a probabilistic topic model. Volume 19. MIT Press (2007) 241–248
18. Potapenko, A.A., Vorontsov, K.V.: Robust PLSA performs better than LDA. In: 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013, Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany (2013) 784–787
19. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476) (2006) 1566–1581
20. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM **57**(2) (2010) 7:1–7:30
21. Tan, Y., Ou, Z.: Topic-weak-correlated latent dirichlet allocation. In: 7th International Symposium Chinese Spoken Language Processing (ISCSLP). (2010) 224–228
22. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: Proceedings of the 24th international conference on Machine learning. ICML '07, New York, NY, USA, ACM (2007) 233–240
23. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the 10th annual Joint Conference on Digital libraries. JCDL '10, New York, NY, USA, ACM (2010) 215–224
24. Newman, D., Bonilla, E.V., Buntine, W.L.: Improving topic coherence with regularized topic models. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K., eds.: Advances in Neural Information Processing Systems 24. (2011) 496–504
25. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 262–272
26. Zhou, S., Li, K., Liu, Y.: Text categorization based on topic model. International Journal of Computational Intelligence Systems **2**(4) (2009) 398–409
27. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. UAI '04, Arlington, Virginia, United States, AUAI Press (2004) 487–494
28. Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., Tong, X.: TextFlow: Towards better understanding of evolving topics in text. IEEE transactions on visualization and computer graphics **17**(12) (2011) 2412–2421
29. Kataria, S., Mitra, P., Caragea, C., Giles, C.L.: Context sensitive topic models for author influence in document networks. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence — Volume 3. IJCAI'11, AAAI Press (2011) 2274–2280
30. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM (2011) 448–456
31. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1) (2010) 1–22

32. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 100–108
33. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/∼mccallum/bow (1996)

Appendix A. The Karush–Kuhn–Tucker (KKT) conditions

Consider the following nonlinear optimization problem:

$$f(x) \to \max_x; \qquad g_i(x) \geq 0,\ \ i = 1,\dots,m; \qquad h_j(x) = 0,\ \ j = 1,\dots,k.$$

Suppose that the objective function $f\colon \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i\colon \mathbb{R}^n \to \mathbb{R}$ and $h_j\colon \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point x^*. If x^* is a local maximum that satisfies some regularity conditions (which are always true if g_i and h_j are linear functions), then there exist constants μ_i, $i = 1,\dots,m$ and λ_j, $j = 1,\dots,k$, called KKT multipliers, such that

$$\frac{\partial}{\partial x}\biggl(f(x) + \sum_{i=1}^{m}\mu_i g_i(x) + \sum_{j=1}^{k}\lambda_j g_j(x)\biggr) = 0; \qquad \text{(stationarity)}$$

$$g_i(x) \geq 0;\ h_j(x) = 0; \qquad \text{(primal feasibility)}$$

$$\mu_i \geq 0; \qquad \text{(dual feasibility)}$$

$$\mu_i g_i(x) = 0. \qquad \text{(complementary slackness)}$$

Appendix B. The Kullback–Leibler divergence

The Kullback–Leibler divergence or relative entropy is a non-symmetric measure of the difference between probability distributions $P = (p_i)_{i=1}^n$ and $Q = (q_i)_{i=1}^n$:

$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^{n} p_i \ln\frac{p_i}{q_i}.$$

From the informational point of view, $\mathrm{KL}(P\|Q)$ is a measure of the information lost when Q is used to approximate P. KL-divergence measures the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P. Typically P represents the empirical distribution of data, Q represents a model or approximation of P.

The KL-divergence is always non-negative.

$\mathrm{KL}(P\|Q) = 0$ if and only if $P = Q$.

The KL-divergence minimization is equivalent to the likelihood maximization of a model distribution $Q(\alpha)$ over parameter vector α:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln\frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad \Longleftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_\alpha.$$