Probable convexity
and its application to Correlated Topic Models[*]

Khoat Than   KHOATTQ@SOICT.HUST.EDU.VN
Hanoi University of Science and Technology, 1 Dai Co Viet road, Hanoi, Vietnam

Tu Bao Ho   BAO@JAIST.AC.JP
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Editor:

Abstract

Non-convex optimization problems often arise from probabilistic modeling, such as estimation of posterior distributions. Non-convexity makes the problems intractable, and poses various obstacles for us to design efficient algorithms. In this work, we attack non-convexity by first introducing the concept of *probable convexity* for analyzing convexity of real functions in practice. We then use the new concept to analyze an inference problem in the *Correlated Topic Model* (CTM) and related nonconjugate models. Contrary to the existing belief of intractability, we show that this inference problem is concave under certain conditions. One consequence of our analyses is a novel algorithm for learning CTM which is significantly more scalable and qualitative than existing methods. Finally, we highlight that stochastic gradient algorithms might be a practical choice to resolve efficiently nonconvex problems. This finding might find beneficial in many contexts which are beyond probabilistic modeling.

Keywords: Non-convex optimization, Posterior estimation, Posterior inference, Nonconjugate models, CTM, Stochastic gradient decent.

1. Introduction

Estimation of posterior distributions plays a central role when developing probabilistic graphical models. With conjugate priors, we are likely able to derive efficient sampling algorithms for estimation (Griffiths and Steyvers, 2004; Pritchard et al., 2000). When nonconjugate priors are used, the estimation problem is much more difficult, as observed in the topic modeling literature by Blei and Lafferty (2007); Salomatin et al. (2009); Putthividhya et al. (2010, 2009); Ahmed and Xing (2007); Blei and Lafferty (2006). A popular approach is to cast estimation as an optimization problem. Nonetheless, the resulting problems are often non-convex. Non-convexity poses various obstacles for designing efficient algorithms, and does not allow us to directly exploit the nice theory of convex optimization.

In this work, we introduce the concept of *probable convexity* that aims at two targets: (1) to reveal how hard an optimization problem in practice is; (2) to support us smoothly employ efficient methods of convex optimization to deal with non-convex problems. In a perspective, probable convexity of a family $\mathfrak{F}$ of real functions essentially says that most members of $\mathfrak{F}$ are convex. With such families, in practice we probably rarely meet non-convex functions

---

[*]. This work was partialy done when K. Than was at JAIST.

from $\mathfrak{F}$. We remark that in many situations of data analytics (e.g., posterior estimation in graphical models) we often have to deal with not only one but many members of a family at once. Hence some appearances of non-convex members may not affect significantly the overall result. Hence a direct employment of convex optimization is possible and beneficial. In other words, we could do minimization efficiently for functions of $\mathfrak{F}$ in practice.

We next use the concept to investigate estimation of posterior distributions in the *Correlated Topic Model* (CTM) (Blei and Lafferty, 2007) and related nonconjugate models. In particular, we study the problem of a posteriori estimating theta (topic mixture) for a given document: $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \Pr(\boldsymbol{\theta}|\boldsymbol{d})$. This is an MAP problem and is intractable for many models in the worst case (Sontag and Roy, 2011). We show that under certain conditions, the objective function of this MAP problem is in fact *probably concave*, i.e., concave with high probability. This suggests that posterior estimation of theta may be tractable in practice. Similar results are obtained for related nonconjugate topic models.

The cornerstone of our analyses of nonconjugate models is the logistic-normal function which originates from the logistic-normal distribution (Aitchison and Shen, 1980). We show in this work that the logistic-normal function is probably concave under certain conditions. This result may be of interest elsewhere and beneficial in practical applications, because the logistic-normal distribution is used as an effective prior in many contexts including topic modeling (Blei and Lafferty, 2007; Salomatin et al., 2009; Putthividhya et al., 2010, 2009; Blei and Lafferty, 2006; Miao et al., 2012) and grammar induction (Cohen and Smith, 2009, 2010).

As a consequence of our analysis, a novel algorithm for learning CTM is proposed. This algorithm is surprisingly simple in which posterior estimation of theta is done by the Online Frank-Wolfe (OFW) algorithm (Hazan and Kale, 2012). From empirical experiments we find that the new algorithm is significantly faster than existing ones, while maintaining or making better the quality of the learned models. This further suggests that even though MAP inference for CTM is intractable in the worst case, most instances in practice may be resolved efficiently.

Finally, we find that stochastic gradient decent (SGD) might be a practical choice to resolve efficiently non-convex problems. SGDs such as OFW (Hazan and Kale, 2012) are originally introduced in the convex optimization literature. They are often very efficient and have many advantages over deterministic algorithms, especially in large-scale settings. However, to our best knowledge, no prior study has been made to investigate the role of SGDs for resolving non-convex problems. We argue that due to their stochastic nature, SGD algorithms might be able to jump out of local optima to reach closer to global ones. Hence SGDs seem to be more advantageous than traditional (deterministic) methods for non-convex problems. We complement this observation by the successful use of OFW to solve posterior estimation of theta in CTM.

Organization: We present the concept of probable convexity in Section 2. Section 3 presents our analysis of the logistic-normal function. The study of CTM and related nonconjugate models is presented in Section 4. The new algorithm for learning CTM and experimental results are discussed in Section 5. We also investigate in this section how well SGDs could resolve non-convex problems by analyzing OFW. The final section is for further discussion and conclusion.

NOTATION: Throughout the paper, we use the following conventions and notations. Bold faces denote vectors or matrices. x_i denotes the ith element of vector $\boldsymbol{x}$, and A_{ij} denotes the element at row i and column j of matrix $\boldsymbol{A}$. Notation $\boldsymbol{A} \leq 0$ means that matrix $\boldsymbol{A}$ is *negative semidefinite*. For a given vector $\boldsymbol{x} = (x_1, ..., x_V)^t$, we denote $\frac{1}{\boldsymbol{x}} = (\frac{1}{x_1}, ..., \frac{1}{x_V})^t$ and $\log \tilde{\boldsymbol{x}} = (\log \frac{x_1}{x_V}, ..., \log \frac{x_{V-1}}{x_V})^t$. $diag(\boldsymbol{x})$ denotes the diagonal matrix whose diagonal entries are $x_1, ..., x_V$, respectively. More notations are:

$\mathcal{V}$:   vocabulary of V terms, often written as $\{1, 2, ..., V\}$.

$\boldsymbol{d}$:   a document represented as a count vector of V dimensions, $\boldsymbol{d} = (d_1, d_2, ..., d_V)$ where d_j is the frequency of term j.

$\mathcal{C}$:   a corpus consisting of M documents, $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_M\}$.

K: number of topics.

$\boldsymbol{\beta}_k$:   a topic which is a distribution over the vocabulary $\mathcal{V}$. It is written as $\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{kV})^t$, where $\beta_{kj} \geq 0, \sum_{j=1}^{V} \beta_{kj} = 1$.

$\mathbb{E}$:   the expectation of a random variable.

Δ_K: the unit simplex in the K-dimensional space, $\Delta_K = \{\boldsymbol{x} \in \mathbb{R}^K : \sum_{k=1}^{K} x_k = 1, x_j \geq 0, \forall j\}$.

$\overline{\Delta}_K$:   the interior of Δ_K, that is $\overline{\Delta}_K = \{\boldsymbol{x} \in \mathbb{R}^K : \sum_{k=1}^{K} x_k = 1, x_j > 0, \forall j\}$.

$\boldsymbol{e}_i$:   the ith unit vector in the Euclidean space, i.e, $e_{ii} = 1$ and $e_{ij} = 0, \forall j \neq i$.

$\exp x$:   denotes e^x.

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:   the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

$x \sim \mathcal{A}(\cdot)$:   the random variable x follows the distribution $\mathcal{A}(\cdot)$.

$\text{Tr}\,\boldsymbol{A}$:   the trace of matrix $\boldsymbol{A}$.

$\lambda_i(\boldsymbol{A})$ the ith largest eigenvalue of matrix $\boldsymbol{A}$.

$\mathbb{S}^K$:   the set of all symmetric matrix of size $K \times K$.

$\mathbb{S}^K_+$:   the set of all positive definite matrices of $\mathbb{S}^K$.

∇f or f':   the gradient (first-order derivative) of the given function f.

f'':   the Hessian matrix (second-order derivative) of the given function f.

$\det \boldsymbol{A}$:   the determinant of the square matrix $\boldsymbol{A}$.

2. Probable convexity

Let $\mathfrak{F}(x; a)$ be a family of real functions defined on a set $X \subset \mathbb{R}^K$, parameterized by a. Each value of a determines a function $f(x; a)$ of $\mathfrak{F}(x; a)$.

Definition 1 (probable convexity) *Let $\mathfrak{F}(x; a)$ be a family of functions defined on a set $X \subset \mathbb{R}^K$, parameterized by a. Family $\mathfrak{F}(x; a)$ is said to be* probably convex *if there exists a positive constant p such that any element of $\mathfrak{F}(x; a)$ is convex on X with probability at least p. Equivalently, $\mathfrak{F}(x; a)$ is said to be* p-convex *if any element of $\mathfrak{F}(x; a)$ is convex on X with probability at least p.*

By definition, a family of convex functions is probably convex with probability 1. The family $\mathfrak{F}(x; a, b, c) = \{ax^2 + bx + c : a, b, c \in \mathbb{R}\}$ is probably convex with probability $1/2$, since convexity of this family is decided by the sign of a.

In a perspective, probable convexity of a family may refer to the proportion of convex members in that family. High p implies that most members are convex on X. Family $\mathfrak{F}(x; a, b, c)$ reflects well this perspective.

In another perspective, p-convexity of $\mathfrak{F}$ may refer to the case that every member of $\mathfrak{F}$ is convex over a part of X. High p implies that the members of $\mathfrak{F}$ is convex over most of X. As an example, family $\mathfrak{F}(x;a) = \{x^4 - 6x^2 + ax : x \in [-10, 10], a \in \mathbb{R}\}$ is 0.9-convex, because each member is convex over 90% of $[-10, 10]$.

Definition 2 (almost sure convexity) *Let $\mathfrak{F}(x;a)$ be a family of functions defined on a set $X \subset \mathbb{R}^K$, parameterized by a. Family $\mathfrak{F}(x;a)$ is said to be* almost surely convex *if any element of $\mathfrak{F}(x;a)$ is convex on X with probability 1.*

It is easy to see that a family of convex functions is almost surely convex. By definition, the family $\mathfrak{F}(x;a,b,c)$ is not almost surely convex. If a family is almost surely convex, almost all of its members are convex.

A family $\mathfrak{F}(x;a)$ is said to be *p-concave* if the family $-\mathfrak{F}(x;a) = \{-f(x;a) : f(x;a) \in \mathfrak{F}(x;a)\}$ is p-convex. One can easily realize that if $\mathfrak{F}(x;a)$ is *p-concave*, then $-\mathfrak{F}(x;a)$ is *p-convex* and vice versa.

The concept of probable convexity applies equally to the cases of only one function. A function $f(x)$ is said to be p-convex in X if it is convex in X with probability at least p. Similarly, function $f(x)$ is said to be p-concave in X if it is concave in X with probability at least p.

Convex optimization refers to minimizing a convex function over a convex domain. It is also refers to maximizing a concave function over a convex domain. It has a long history and has a rich foundation. Convex problems are often considered as being easy since there exist various fast algorithms. The book by Boyd and Vandenberghe (2004) provides an excellent introduction to the field.

3. Concavity of the logistic-normal function

We first consider probable convexity of the following function which is called *logistic-normal*:

$$LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu}) - \sum_{k=1}^{K} \log x_k, \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{K-1}, \boldsymbol{\Sigma} \in \mathbb{S}_+^{K-1}; \boldsymbol{x} \in \overline{\Delta}_K$ such that $\log \tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This function naturally originates from the logistic-normal distribution (Aitchison and Shen, 1980), whose density is $p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp(LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))$. Due to the broad use of this distribution in probabilistic modeling, the logistic-normal function plays an important role in many contexts. Nonetheless, the function itself is neither convex nor concave in $\overline{\Delta}_K$. This is one of the main reasons for why posterior estimation in nonconjugate models is often intractable.

By a thorough analysis of this function, we found the following property.

Theorem 1 *Denote $p = 1 - e^{2\log(K-1) - 0.5(\lambda-1)^2/\sigma}$ for $\lambda = \lambda_{K-1}(\boldsymbol{\Sigma}^{-1})$ and $\sigma = \max_i \Sigma_{ii}^{-1}$. Function $LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is p-concave over $\overline{\Delta}_K$ if $\lambda \geq 1$.*

This theorem essentially says that LN is in fact concave under some conditions. Note that the quantity $(\lambda - 1)^2/\sigma$ is not always small. Indeed, letting $\lambda_k(\boldsymbol{\Sigma}^{-1})$ be the kth eigenvalue of $\boldsymbol{\Sigma}^{-1}$, we have $\text{Tr}(\boldsymbol{\Sigma}^{-1}) = \sum_{k=1}^{K-1} \lambda_k(\boldsymbol{\Sigma}^{-1}) = \sum_{k=1}^{K-1} \Sigma_{kk}^{-1}$. When the condition

number of $\mathbf{\Sigma}^{-1}$ is not large, $\lambda_{K-1}(\mathbf{\Sigma}^{-1})$ and σ may be of the same order. This observation suggests that the probability bound obtained in Theorem 1 is significant.

Corollary 1 *With notations as in Theorem 1, function $LN(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is almost surely concave as $\lambda^2/\sigma \to +\infty$.*

In the case that the least eigenvalue λ is much larger than $\log(K-1)$, function LN is concave with high probability. More concretely, if $\lambda^2 = \omega(\sigma \log K)$, i.e., $\lambda^2/\sigma \log K \to +\infty$ as $K \to +\infty$, then $\exp\left\{2\log(K-1) - 0.5(\lambda-1)^2/\sigma\right\}$ goes to 0. Hence the following result holds.

Corollary 2 *With notations as in Theorem 1, assume that $\lambda^2 = \omega(\sigma \log K)$. Function $LN(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ is almost surely concave as $K \to +\infty$.*

3.1 Proof of Theorem 1

We will show probable concavity of LN by investigating concavity in common sense. Note that the domain $\overline{\Delta}_K$ is convex, and function LN is twice differentiable over $\overline{\Delta}_K$. Hence, to see concavity, it suffices to show that the second derivative is negative semidefinite (Boyd and Vandenberghe, 2004).

Let Σ_i^{-1} be the ith row of $\mathbf{\Sigma}^{-1}$. The first and second partial derivatives of the function w.r.t the variables are:

$$
\frac{\partial LN}{\partial x_i} = \begin{cases} -\frac{1}{x_i}\Sigma_i^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu}) - \frac{1}{x_i}, & i < K \\ \frac{1}{x_K}\sum_{h=1}^{K-1}\Sigma_h^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu}) - \frac{1}{x_K}, & i = K \end{cases}
$$

$$
\frac{\partial^2 LN}{\partial x_i \partial x_j} = \begin{cases} -\frac{\Sigma_{ij}^{-1}}{x_i x_j}, & i < K, i \neq j, j < K \\ \frac{1}{x_i^2}\Sigma_i^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu}) - \frac{\Sigma_{ii}^{-1}}{x_i^2} + \frac{1}{x_i^2}, & i < K, i = j \\ \frac{1}{x_i x_K}\sum_{h=1}^{K-1}\Sigma_{ih}^{-1}, & i < K, j = K \\ \frac{1}{x_j x_K}\sum_{h=1}^{K-1}\Sigma_{hj}^{-1}, & i = K, j < K \\ -\frac{1}{x_K^2}\sum_{h=1}^{K-1}\Sigma_h^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu}) - \frac{1}{x_K^2}\sum_{h=1}^{K-1}\sum_{t=1}^{K-1}\Sigma_{ht}^{-1} + \frac{1}{x_K^2}, & i = j = K. \end{cases}
$$

Denote $\boldsymbol{S} = \begin{pmatrix} \mathbf{\Sigma}^{-1} & \boldsymbol{s}_K^t \\ \boldsymbol{s}_K & s_{KK} \end{pmatrix}; \boldsymbol{U} = \begin{pmatrix} \mathbf{\Sigma}^{-1} \\ \boldsymbol{s}_K \end{pmatrix}$, where $\boldsymbol{s}_K = -\sum_{t=1}^{K-1}\Sigma_t^{-1}$ is the sum of the rows of $\mathbf{\Sigma}^{-1}$, and s_{KK} is the sum of all elements of $\mathbf{\Sigma}^{-1}$. We can express the second derivative of LN as

$$
\begin{aligned}
LN'' &= diag\frac{1}{\boldsymbol{x}}.diag[\boldsymbol{U}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})].diag\frac{1}{\boldsymbol{x}} - diag\frac{1}{\boldsymbol{x}}.\boldsymbol{S}.diag\frac{1}{\boldsymbol{x}} + diag\frac{1}{\boldsymbol{x}}.diag\frac{1}{\boldsymbol{x}} \\
&= diag\frac{1}{\boldsymbol{x}}.\left(\boldsymbol{I}_K - \boldsymbol{S} + diag[\boldsymbol{U}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})]\right).diag\frac{1}{\boldsymbol{x}}.
\end{aligned} \tag{2}
$$

A classical result in Algebra (Abadir and Magnus, 2005, exercise 8.28) says that for any symmetric $\boldsymbol{A}$ and nonsingular $\boldsymbol{Y}$, the product $\boldsymbol{Y}\boldsymbol{A}\boldsymbol{Y}^t$ is positive semidefinite if and only if $\boldsymbol{A}$ is positive semidefinite. Consequently, the matrix $\boldsymbol{I}_K - \boldsymbol{S} + diag[\boldsymbol{U}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})]$ decides negative semidefiniteness of LN''.

Lemma 1 *Denote $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})$. LN'' is negative semidefinite if $z_1 + \cdots + z_{K-1} \geq 1$ and $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z}) \leq 0$.*

Proof As discussed before, matrix $\boldsymbol{I}_K - \boldsymbol{S} + diag[\boldsymbol{U}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})]$ decides negative semidefiniteness of LN''. Letting $z_K = -z_1 - \cdots - z_{K-1}$ and $\boldsymbol{1} = (1, ..., 1)^t \in \mathbb{R}^{K-1}$, we have

$$
\begin{aligned}
\boldsymbol{A} &= \boldsymbol{I}_K - \boldsymbol{S} + diag[\boldsymbol{U}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})] \\
&= \boldsymbol{I}_K - (\boldsymbol{I}_{K-1} \; \boldsymbol{1})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{I}_{K-1} \; \boldsymbol{1}) + diag(z_1, ..., z_K) \\
&= (\boldsymbol{I}_{K-1} \; \boldsymbol{1})^t \left[\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z})\right](\boldsymbol{I}_{K-1} \; \boldsymbol{1}) + \begin{pmatrix} \boldsymbol{0} & -(\boldsymbol{z}+\boldsymbol{1}) \\ -(\boldsymbol{z}+\boldsymbol{1})^t & z_K + 1 \end{pmatrix} \quad (3)
\end{aligned}
$$

Consider the last term $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{0} & -(\boldsymbol{z}+\boldsymbol{1}) \\ -(\boldsymbol{z}+\boldsymbol{1})^t & z_K + 1 \end{pmatrix}$. This matrix is of size $K \times K$, but has rank 2. It is not hard to see that all principle minors of $\boldsymbol{C}$ are 0, except the ones which associate with the last two rows and columns. Those principle minors are $z_K + 1$ and $\begin{vmatrix} 0 & -z_i - 1 \\ -z_i - 1 & z_K + 1 \end{vmatrix} = z_K + 1 - (z_i + 1)^2$ for $i \in \{1, ..., K-1\}$. According to a classical result in Algebra (Abadir and Magnus, 2005, exercise 8.32), $\boldsymbol{C} \leq 0$ if and only if all of its principle minors are non-positive. Therefore $\boldsymbol{C} \leq 0$ if and only if $z_K + 1 \leq 0$.

If $\boldsymbol{C}$ and $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z})$ are negative semidefinite, so are $\boldsymbol{A}$ and LN''. This suggests that if $z_K + 1 \leq 0$ and $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z}) \leq 0$, then $LN'' \leq 0$ which completes the proof. ∎

Next we want to see under what conditions, matrix $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z}) \leq 0$ with the constraint of $z_1 + \cdots + z_{K-1} \geq 1$. The following theorem reveals a property whose detailed proof is presented in section 3.2.

Theorem 2 *Let $\boldsymbol{z}$ be a Gaussian random variable with mean 0 and covariance matrix $\boldsymbol{A} \in \mathbb{S}_+^{K-1}$, and $\sigma = \max_i A_{ii}$. For a fixed $\boldsymbol{S} \in \mathbb{S}_+^{K-1}$, consider $\boldsymbol{B} = \boldsymbol{I}_{K-1} - \boldsymbol{S} + diag(\boldsymbol{z})$. Assuming $\lambda_{K-1}(\boldsymbol{S}) \geq 1$, we have*

$$
\Pr(\lambda_1(\boldsymbol{B}) \geq 0 | z_1 + \cdots + z_{K-1} \geq 1) \leq \exp\left\{2\log(K-1) - 0.5(1 - \lambda_{K-1}(\boldsymbol{S}))^2/\sigma\right\}.
$$

This theorem essentially says that under certain assumption, matrix $\boldsymbol{B}$ is negative semidefinite with probability at least $1 - \exp\left\{2\log(K-1) - 0.5(1 - \lambda_{K-1}(\boldsymbol{S}))^2/\sigma\right\}$. Hence we have enough tools to prove Theorem 1.

Proof [Proof of Theorem 1] Consider the logistic-normal function $LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and denote $\lambda = \lambda_{K-1}(\boldsymbol{\Sigma}^{-1})$ and $\sigma = \max_i \Sigma_{ii}^{-1}$. As discussed before, concavity of this function over $\overline{\Delta}_K$ is decided by its second partial derivative LN''. Lemma 1 suggests that $LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is concave if $z_1 + \cdots + z_{K-1} \geq 1$ and $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z}) \leq 0$, where $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1}(\log \tilde{\boldsymbol{x}} - \boldsymbol{\mu})$. Note that $\mathbb{E}\boldsymbol{z} = 0$ and $cov(z) = \boldsymbol{\Sigma}^{-1}$ since $\mathbb{E} \log \tilde{\boldsymbol{x}} = \boldsymbol{\mu}$ and $cov(\log \tilde{\boldsymbol{x}}) = \boldsymbol{\Sigma}$. Theorem 2 implies that with the constraint of $z_1 + \cdots + z_{K-1} \geq 1$, $\boldsymbol{I}_{K-1} - \boldsymbol{\Sigma}^{-1} + diag(\boldsymbol{z}) \leq 0$ holds with probability at least $1 - \exp\left\{2\log(K-1) - 0.5(1 - \lambda)^2/\sigma\right\}$ if $\lambda \geq 1$. This means assuming $\lambda \geq 1$, function $LN(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is concave with probability at least $1 - \exp\left\{2\log(K-1) - 0.5(1 - \lambda)^2/\sigma\right\}$. ∎

3.2 Proof of Theorem 2

To prove this theorem we need some basic results from matrix algebra and the theory of random matrices.

A matrix $\boldsymbol{A}$ is *positive semidefinite* if and only if the least eigenvalue $\lambda_{\min}(\boldsymbol{A})$ is non-negative. If $\boldsymbol{A}$ has K eigenvalues, its trace satisfies $\operatorname{Tr}\boldsymbol{A} = \sum_{i=1}^{K} \lambda_i(\boldsymbol{A})$. If $\boldsymbol{A}$ is a random matrix, we have trace-expectation relation $\operatorname{Tr}\mathbb{E}\boldsymbol{A} = \mathbb{E}(\operatorname{Tr}\boldsymbol{A})$.

Consider a function $f : \mathbb{R} \to \mathbb{R}$. We define a map on a diagonal matrix $\boldsymbol{A} \in \mathbb{S}^K$ as $f(\boldsymbol{A}) = diag(f(A_{11}), ..., f(A_{KK}))$. Similarly, a function of a symmetric matrix $\boldsymbol{A}$ is defined by using the eigenvalue decomposition:

$$f(\boldsymbol{A}) = \boldsymbol{Q}.f(\boldsymbol{\Lambda}).\boldsymbol{Q}^t, \text{ where } \boldsymbol{A} = \boldsymbol{Q}.\boldsymbol{\Lambda}.\boldsymbol{Q}^t \text{ and } \boldsymbol{\Lambda} \text{ is a diagonal matrix.}$$

The ***spectral mapping theorem*** states that each eigenvalue of $f(\boldsymbol{A})$ is equal to $f(\lambda)$ for some eigenvalue λ of $\boldsymbol{A}$. If f is nondecreasing, then $\lambda_k(f(\boldsymbol{A})) = f(\lambda_k(\boldsymbol{A}))$ for any k whenever $\lambda_k(\boldsymbol{A})$ exists.

We will work with *matrix exponential* which is defined for an $\boldsymbol{A} \in \mathbb{S}^K$ by

$$e^{\boldsymbol{A}} = \sum_{i=0}^{\infty} \frac{\boldsymbol{A}^i}{i!}.$$

Note that $\lambda_k(e^{\boldsymbol{A}}) = e^{\lambda_k(\boldsymbol{A})}$ for any k provided that $\lambda_k(\boldsymbol{A})$ exists. The logarithm of a matrix $\boldsymbol{A} \in \mathbb{S}_+^K$ is a matrix, denoted by $\log\boldsymbol{A}$, such that $e^{\log\boldsymbol{A}} = \boldsymbol{A}$.

Theorem 3 (Golden-Thompson inequality) *For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^K$, we have*

$$\operatorname{Tr} e^{\boldsymbol{A}+\boldsymbol{B}} \leq \operatorname{Tr}(e^{\boldsymbol{A}}.e^{\boldsymbol{B}}).$$

This is a standard result and can be found in (Wigderson and Xiao, 2008; Tropp, 2012). Note that $e^{\boldsymbol{A}}$ and $e^{\boldsymbol{B}}$ are positive definite which implies $\operatorname{Tr}(e^{\boldsymbol{A}}.e^{\boldsymbol{B}}) \leq \operatorname{Tr} e^{\boldsymbol{A}}.\operatorname{Tr} e^{\boldsymbol{B}}$, since according to Yang and Feng (2002), $\operatorname{Tr}(\boldsymbol{A}.\boldsymbol{B}) \leq \operatorname{Tr}\boldsymbol{A}.\operatorname{Tr}\boldsymbol{B}$ if $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}_+^K$. Hence we have the following.

Corollary 3 *For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^K$, we have $\operatorname{Tr} e^{\boldsymbol{A}+\boldsymbol{B}} \leq \operatorname{Tr} e^{\boldsymbol{A}}.\operatorname{Tr} e^{\boldsymbol{B}}$.*

The next theorem was shown by Tropp (2012).

Theorem 4 (Laplace transform method) *Let $\boldsymbol{B}$ be a random matrix of $\mathbb{S}^K$. For any real t, we have*

$$\Pr(\lambda_1(\boldsymbol{B}) \geq t) \leq \inf_{a>0}\{e^{a.t}\mathbb{E}\operatorname{Tr} e^{a.\boldsymbol{B}}\}.$$

Lemma 2 *Consider a matrix $\boldsymbol{B} \in \mathbb{S}^K$ and a nonnegative real a. We have*

$$\mathbb{E}\operatorname{Tr} e^{a.\boldsymbol{B}} \leq K\mathbb{E}e^{a\lambda_1(\boldsymbol{B})}.$$

Proof Since the trace of $\boldsymbol{B}$ equals the sum of its eigenvalues, we have $\operatorname{Tr}\boldsymbol{B} \leq K\lambda_1(\boldsymbol{B})$. Hence $\mathbb{E}\operatorname{Tr} e^{a.\boldsymbol{B}} \leq K\mathbb{E}\lambda_1(e^{a\boldsymbol{B}}) \leq K\mathbb{E}e^{\lambda_1(a\boldsymbol{B})} = K\mathbb{E}e^{a\lambda_1(\boldsymbol{B})}$, where the last inequality is derived by using the spectral mapping theorem. ∎

Lemma 3 *Consider a Gaussian random vector $\boldsymbol{z}$ with mean 0 and covariance matrix $\boldsymbol{A} \in \mathbb{S}_+^K$. Let $\sigma_i = A_{ii}$ be the ith diagonal entry of $\boldsymbol{A}$, and $\sigma = \max_i \sigma_i$. Then for any real $a > 0$, we have $\mathbb{E} \operatorname{Tr} e^{a.diag(\boldsymbol{z})} = \sum_{k=1}^K e^{a^2 \sigma_k/2} \le K e^{a^2 \sigma/2}$.*

Proof Note that

$$
\begin{aligned}
\operatorname{Tr} e^{a.diag(\boldsymbol{z})} &= \operatorname{Tr} \sum_{i=0}^{\infty} \frac{a^i}{i!} diag^i(\boldsymbol{z}) \\
&= \operatorname{Tr} \sum_{i=0}^{\infty} \frac{a^i}{i!} diag(z_1^i, ..., z_K^i) \\
&= \sum_{i=0}^{\infty} \frac{a^i}{i!} \operatorname{Tr} diag(z_1^i, ..., z_K^i) \\
&= \sum_{i=0}^{\infty} \frac{a^i}{i!} \sum_{k=1}^K z_k^i = \sum_{k=1}^K \sum_{i=0}^{\infty} \frac{a^i}{i!} z_k^i = \sum_{k=1}^K e^{az_k}
\end{aligned}
$$

Hence $\mathbb{E} \operatorname{Tr} e^{a.diag(\boldsymbol{z})} = \mathbb{E} \sum_{k=1}^K e^{a.z_k} = \sum_{k=1}^K \mathbb{E} e^{a.z_k}$.

By assumption, z_k is a Gaussian variable with mean 0 and variance σ_k. Using the generating function of Gaussian, we have $\mathbb{E} e^{a.z_k} = e^{a^2 \sigma_k/2}$. So substituting these quantities into the expectation in the last paragraph completes the proof. ∎

Proof [Proof of Theorem 2]

We have

$$
\begin{aligned}
\Pr(\lambda_1(\boldsymbol{B}) \ge 0 | z_1 + \cdots + z_{K-1} \ge 1) &\le \Pr(\lambda_1(\boldsymbol{B}) \ge 0) \\
&\le \inf_{a>0} \left\{ \mathbb{E} \operatorname{Tr} e^{a\boldsymbol{B}} \right\} \\
&\quad \text{(Laplace transform method)} \\
&= \inf_{a>0} \left\{ \mathbb{E} \operatorname{Tr} e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S} + diag(\boldsymbol{z})]} \right\} \\
&\le \inf_{a>0} \left\{ \mathbb{E} \left(\operatorname{Tr} e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S}]} . \operatorname{Tr} e^{a.diag(\boldsymbol{z})} \right) \right\} \\
&\quad \text{(Corollary 3)} \\
&= \inf_{a>0} \left\{ \operatorname{Tr} e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S}]} . \mathbb{E} \operatorname{Tr} e^{a.diag(\boldsymbol{z})} \right\} \\
&\le \inf_{a>0} \left\{ \operatorname{Tr} e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S}]} . (K-1) . e^{a^2 \sigma/2} \right\} \\
&\quad \text{(Lemma 3)} \\
&= \inf_{a>0} \left\{ (K-1) . e^{a^2 \sigma/2} . \operatorname{Tr} e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S}]} \right\} \\
&\le \inf_{a>0} \left\{ (K-1) . e^{a^2 \sigma/2} . (K-1) . \lambda_1(e^{a[\boldsymbol{I}_{K-1} - \boldsymbol{S}]}) \right\}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(\lambda_1(\boldsymbol{B}) \geq 0 | z_1 + \cdots + z_{K-1} \geq 1) \ &\leq\ \inf_{a>0} \left\{ (K-1)^2 . e^{a^2\sigma/2} . e^{\lambda_1(a[\boldsymbol{I}_{K-1}-\boldsymbol{S}])} \right\} \\
&\qquad \text{(Spectral mapping theorem)} \\
&=\ \inf_{a>0} \left\{ (K-1)^2 . e^{a^2\sigma/2} . e^{a-a\lambda_{K-1}(\boldsymbol{S})} \right\} \\
&=\ \inf_{a>0} \left\{ (K-1)^2 . e^{a^2\sigma/2 + a - a\lambda_{K-1}(\boldsymbol{S})} \right\} \\
&=\ (K-1)^2 \exp \left\{ -\frac{(1-\lambda_{K-1}(\boldsymbol{S}))^2}{2\sigma} \right\}.
\end{aligned}
$$

Note that the last equality is obtained by minimizing the function $a^2\frac{\sigma}{2} + a - a\lambda_{K-1}(\boldsymbol{S})$ for $a > 0$ conditioned on $1 \leq \lambda_{K-1}(\boldsymbol{S})$. ∎

4. MAP inference of topic mixtures in CTM

We next study convexity of a family originated from the topic modeling literature. In particular, we are interested in the problem of estimating topic mixtures (posterior distributions) in correlated topic models (CTM) (Blei and Lafferty, 2007). This problem is intractable by traditional approaches (Blei and Lafferty, 2007; Ahmed and Xing, 2007). We will show that in fact this problem is tractable under some conditions, by showing probable concavity.

The correlated topic model assumes that a corpus is composed from K topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$, and a document $\boldsymbol{d}$ arises from the following generative process:

1. Draw $\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

2. For the nth word of $\boldsymbol{d}$:

  - draw topic assignment $z_{dn} | \boldsymbol{x} \sim \mathcal{M}(f(\boldsymbol{x}))$
  - draw word $w_{dn} | z_{dn}, \boldsymbol{\beta} \sim \mathcal{M}(\boldsymbol{\beta}_{z_{dn}})$.

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathcal{M}(\cdot)$ is the multinomial distribution; $f(\boldsymbol{x})$ maps a natural parameterization of the topic proportion to the mean parameterization:

$$
\boldsymbol{\theta} = f(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}}}{\sum_{k=1}^{K} e^{x_k}}. \tag{4}
$$

This logistic transformation maps a K-dimensional vector $\boldsymbol{x}$ to a $(K-1)$-dimensional vector $\boldsymbol{\theta}$. Hence various $\boldsymbol{x}$'s can correspond to a single $\boldsymbol{\theta}$. Fixing $x_K = 0$, the transformation (4) means that $\boldsymbol{\theta}$ follows the logistic-normal distribution (Blei and Lafferty, 2007). According to Aitchison and Shen (1980), the density function of $\boldsymbol{\theta}$ is thus

$$
p(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp \left(-\frac{1}{2} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^{K} \log \theta_k \right), \tag{5}
$$

where $\boldsymbol{\mu} \in \mathbb{R}^{K-1}, \boldsymbol{\Sigma} \in \mathbb{S}_+^{K-1}$. Note that $\boldsymbol{\theta}$ is derived from $\boldsymbol{x}$ by (4). Hence $\log \tilde{\boldsymbol{\theta}}$ is a normal random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

One of the most interesting tasks in this model is the posterior estimation of topic mixtures for documents. More concretely, given the model parameters $\Upsilon = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, we are interested in the following problem for a given document $\boldsymbol{d}$:

$$
\begin{aligned}
\boldsymbol{\theta}^* &= \arg \max_{\boldsymbol{\theta} \in \Delta_K} \Pr(\boldsymbol{\theta} | \boldsymbol{d}, \Upsilon) \\
&= \arg \max_{\boldsymbol{\theta} \in \Delta_K} \Pr(\boldsymbol{\theta}, \boldsymbol{d} | \Upsilon) \quad (6)
\end{aligned}
$$

Lemma 4 *Given a CTM model with parameters* $\Upsilon = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ *and a document* $\boldsymbol{d}$*, the MAP problem (6) can be reformulated as*

$$
\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \overline{\Delta}_K} \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} - \frac{1}{2} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^K \log \theta_k. \quad (7)
$$

Proof We have

$$
\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \Delta_K} \Pr(\boldsymbol{\theta}, \boldsymbol{d} | \Upsilon) = \arg \max_{\boldsymbol{\theta} \in \Delta_K} \log \Pr(\boldsymbol{\theta}, \boldsymbol{d} | \Upsilon) = \arg \max_{\boldsymbol{\theta} \in \Delta_K} \log \Pr(\boldsymbol{d} | \boldsymbol{\theta}, \Upsilon) + \log \Pr(\boldsymbol{\theta} | \Upsilon).
$$

Note that $\log \Pr(\boldsymbol{d} | \boldsymbol{\theta}, \Upsilon) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$, and the density of the logistic-normal distribution is given in (5). Hence

$$
\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \Delta_K} \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} - \frac{1}{2} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^K \log \theta_k - \frac{1}{2} \log \det(2\pi \boldsymbol{\Sigma}).
$$

Since any point on the boundary of Δ_K makes the objective function undefined and hence is not optimal. Therefore, ignoring the boundary of Δ_K and the constant in the objective function completes the proof. ■

Loosely speaking, Lemma 4 says that posterior estimation of topic mixtures in CTM is in fact an optimization problem. The objective function is well-defined on $\overline{\Delta}_K$. It is worth remarking that this function is neither concave nor convex in general. Hence maximizing it over $\overline{\Delta}_K$ is intractable in the worse case.

4.1 Some results

Let the model parameters $\Upsilon = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ be fixed, where $\boldsymbol{\beta}_k \in \Delta_V, \boldsymbol{\mu} \in \mathbb{R}^{K-1}, \boldsymbol{\Sigma} \in \mathbb{S}_+^{K-1}$. Consider the following family, parameterized by $\boldsymbol{d}$:

$$
CTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon) = \{f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon) : \boldsymbol{\theta} \in \overline{\Delta}_K, \log \tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\}. \quad (8)
$$

where $f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} - \frac{1}{2} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^K \log \theta_k$. This family contains all possible instances of the problem (7). Hence, analyzing this family means analyzing the problem of estimating topic mixtures in CTM.

Consider a member $f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$. Note that $\boldsymbol{d}$ and $\boldsymbol{\beta}$ are always nonnegative in practices of topic modeling. Hence the first term in $f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is always concave over $\overline{\Delta}_K$. It implies that concavity of $f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is heavily determined by the logistic-normal term $y = -\frac{1}{2} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^K \log \theta_k$. If this term is concave, then $f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is concave. Combining these observations with Theorem 1, Corollary 1, and Corollary 2, we arrive at the following results for CTM.

Theorem 5 *Let Υ be fixed, $\sigma = \max_i \Sigma_{ii}^{-1}, \lambda = \lambda_{K-1}(\mathbf{\Sigma}^{-1})$, and $p = 1 - e^{2\log(K-1)-0.5(\lambda-1)^2/\sigma}$. Assuming $\lambda \geq 1$, family $CTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is p-concave over $\overline{\Delta}_K$.*

Corollary 4 *With notations as in Theorem 5, family $CTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is almost surely concave as $\lambda^2/\sigma \to +\infty$.*

Corollary 5 *With notations as in Theorem 5, assume that $\lambda^2 = \omega(\sigma \log K)$. Family $CTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon)$ is almost surely concave as $K \to +\infty$.*

4.2 Implication to related models

Many nonconjugate models employ the Gaussian distribution to model correlation of hidden topics, including those by Blei and Lafferty (2006); Putthividhya et al. (2009, 2010); Salomatin et al. (2009); Miao et al. (2012). The analysis for CTM is very general for the case of logistic-normal priors. Therefore, the results for CTM can be easily derived for other nonconjugate topic models. Here we take DTM (Blei and Lafferty, 2006) and IFTM (Putthividhya et al., 2009) into consideration as two specific examples.

The *Independent Factor Topic Model* (IFTM) by Putthividhya et al. (2009) is a variant of CTM in which $\boldsymbol{\mu}$ is replaced with $\boldsymbol{\mu}' = \boldsymbol{As} + \boldsymbol{\mu}$ to model independent sources that compose correlated topics. A slight modification to our analysis would yield interesting results for the corresponding family, denoting $\Upsilon' = \{\boldsymbol{\beta}, \boldsymbol{\mu}', \boldsymbol{\Sigma}\}$,

$$IFTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon') = \{f(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon') : \boldsymbol{\theta} \in \overline{\Delta}_K, \log \tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma})\}.$$

Theorem 6 *Let Υ' be fixed, $\sigma = \max_i \Sigma_{ii}^{-1}, \lambda = \lambda_{K-1}(\mathbf{\Sigma}^{-1})$, and $p = 1 - e^{2\log(K-1)-0.5(\lambda-1)^2/\sigma}$. Assuming $\lambda \geq 1$, family $IFTM(\boldsymbol{\theta}; \boldsymbol{d}, \Upsilon')$ is p-concave over $\overline{\Delta}_K$.*

The *Dynamic Topic Model* (DTM) by Blei and Lafferty (2006) also employs Gaussian priors to model correlation. Those priors are separable, i.e., having diagonal covariance matrices. Let $DTM(\boldsymbol{\theta}; \boldsymbol{d}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma)$ be defined similarly with (8), where $\mathbf{\Sigma}^{-1} = diag(\sigma, ..., \sigma)$. For this family, note that $\lambda_{K-1}(\mathbf{\Sigma}^{-1}) = \sigma$. Hence, Theorem 5 implies

Theorem 7 *For fixed $\{\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma\}$, if $\sigma \geq 1$ then family $DTM(\boldsymbol{\theta}; \boldsymbol{d}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma)$ is probably concave with probability at least $1 - e^{2\log(K-1)-0.5\sigma-0.5/\sigma+1}$.*

5. A fast algorithm for learning CTM

In this section we discuss an application of the findings in Section 4 to designing an efficient algorithm for learning CTM. Nonconjugacy of the prior over $\boldsymbol{\theta}$ poses various drawbacks and precludes using sampling techniques. Hence Blei and Lafferty (2007) proposed to use variational Bayesian methods to approximate the posterior distributions of latent variables. Variational Bayesian methods have been employed heavily for learning many other nonconjugate models (Salomatin et al., 2009; Putthividhya et al., 2010, 2009; Blei and Lafferty, 2006; Miao et al., 2012). The use of simplified distributions to approximate the true posterior often results in more parameters to be optimized when learning a model. (For example, the method by Blei and Lafferty (2007) maintains K Gaussian distributions for each document.) Hence it could be problematic when the corpus is large.

Learning CTM and other related models can be made significantly simpler by using our analysis. Indeed, to estimate the posterior ($P(\boldsymbol{\theta}|\boldsymbol{d}, \Upsilon)$) of topic mixtures, one can exploit fast algorithms for convex optimization. The analysis in Section 4 provides a theoretically reasonable justification for such an exploitation. Once $\boldsymbol{\theta}$ had been inferred for each document in the training data, one can follow the approach by Than and Ho (2012) to estimate topics $\boldsymbol{\beta}$. A Gaussian prior is also easily estimated when all $\boldsymbol{\theta}$ of the training documents are known.

5.1 Derivation of the algorithm

Our proposed algorithm for learning CTM is presented in Algorithm 1 which is an alternative algorithm similar to EM. This algorithm tries to maximize the following regularized joint likelihood of the training corpus $\mathcal{C}$:

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{\boldsymbol{d} \in \mathcal{C}} \log \Pr(\boldsymbol{\theta}, \boldsymbol{d}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \frac{M}{2}\alpha \operatorname{Tr} \boldsymbol{\Sigma}^{-1} \\
&= \sum_{\boldsymbol{d} \in \mathcal{C}} \sum_{j} d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} - \frac{1}{2} \sum_{\boldsymbol{d} \in \mathcal{C}} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) \\
&\quad - \frac{M}{2} \log \det \boldsymbol{\Sigma} - \frac{M}{2}\alpha \operatorname{Tr} \boldsymbol{\Sigma}^{-1} + constant.
\end{aligned}
$$

The main reason for imposing a regularization term $\alpha \operatorname{Tr} \boldsymbol{\Sigma}^{-1}$ on the joint likelihood is to control the eigenvalues of the learned $\boldsymbol{\Sigma}^{-1}$. Large α often prevents the eigenvalues of $\boldsymbol{\Sigma}^{-1}$ from increasing. On the other hand, small values of α play the role as allowing large eigenvalues of $\boldsymbol{\Sigma}^{-1}$. In the latter case, Corollary 4 and Corollary 5 suggest that estimation of topic mixtures ($\boldsymbol{\theta}$) is more likely to be a concave problem, and thus can be done efficiently.

In Step 1 which does posterior inference for each document, we use the Online Frank-Wolfe algorithm (Hazan and Kale, 2012) to maximize the joint probability $\Pr(\boldsymbol{\theta}, \boldsymbol{d}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. This algorithm theoretically converges to the optimal solutions, provided that the optimization problem is concave.[1] Note that Algorithm 2 is a slight but careful modification of the general algorithm by Hazan and Kale (2012), and in fact is similar with the algorithm which is presented by Clarkson (2010).

In Step 2, we fix $\boldsymbol{\theta}_d$ which has been inferred for each document $\boldsymbol{d} \in \mathcal{C}$ in Step 1, and maximize $L(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to estimate the model parameters. Solving for $\boldsymbol{\beta}$ can be done independently of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Hence by using the same argument as Than and Ho (2012), we can arrive at the formula (10) for updating topics. Maximizing the term relating to $\boldsymbol{\mu}$ in $L(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ will lead to (11) for updating $\boldsymbol{\mu}$.

Take $\boldsymbol{\Sigma}$ into consideration: $L_\alpha = -\frac{1}{2} \sum_{\boldsymbol{d} \in \mathcal{C}} (\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu}) - \frac{M}{2} \log \det \boldsymbol{\Sigma} - \frac{M}{2}\alpha \operatorname{Tr} \boldsymbol{\Sigma}^{-1}$. Its derivative with respect to $\boldsymbol{\Sigma}^{-1}$ is $\nabla L_\alpha = -\frac{1}{2} \sum_{\boldsymbol{d} \in \mathcal{C}} (\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu})(\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu})^t + \frac{M}{2}\boldsymbol{\Sigma} - \frac{M}{2}\alpha \boldsymbol{I}_{K-1}$. Solving $\nabla L_\alpha = 0$, one can derive (12) for updating $\boldsymbol{\Sigma}$.

---

1. In practice we can approximate $\overline{\Delta}_K$ by $\Delta_\epsilon = \{\boldsymbol{\theta} : \sum_{k=1}^{K} \theta_k = 1, \theta_i \geq \epsilon, \forall i\}$ for a very small constant ϵ, says $\epsilon = 10^{-10}$. Hence the online Frank-Wolfe algorithm should be slightly modified accordingly.

Algorithm 1 fCTM: a fast algorithm for learning correlated topic models

---

Input: a corpus $\mathcal{C} = \{\boldsymbol{d_1}, ..., \boldsymbol{d_M}\}$, and a positive constant α.

Output: $\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

Initialize $\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$, and then alternate the following two steps until convergence.

Step 1: for each document $\boldsymbol{d}$, use Algorithm 2 to solve for

$$\boldsymbol{\theta}_d = \arg \max_{\boldsymbol{\theta} \in \overline{\Delta}_K} \log \Pr(\boldsymbol{\theta}, \boldsymbol{d}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

Step 2: compute

$$\beta_{kj} \quad \propto \quad \sum_{\boldsymbol{d} \in \mathcal{C}} d_j \theta_{dk}, \tag{10}$$

$$\boldsymbol{\mu} \quad = \quad \frac{1}{M} \sum_{\boldsymbol{d} \in \mathcal{C}} \log \tilde{\boldsymbol{\theta}}_d, \tag{11}$$

$$\boldsymbol{\Sigma} \quad = \quad \alpha \boldsymbol{I}_{K-1} + \frac{1}{M} \sum_{\boldsymbol{d} \in \mathcal{C}} (\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu})(\log \tilde{\boldsymbol{\theta}}_d - \boldsymbol{\mu})^t. \tag{12}$$

Algorithm 2 Online Frank-Wolfe (OFW)

Input: document $\boldsymbol{d}$, and model $\Upsilon = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

Output: $\boldsymbol{\theta}$ that maximizes
$f(\boldsymbol{\theta}) = \sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} - \frac{1}{2}(\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^{K} \log \theta_k$.

Initialize $\boldsymbol{\theta}_1$ arbitrarily in $\overline{\Delta}_K$.

for $\ell = 1, ..., \infty$ **do**

    Pick f_ℓ uniformly from
    $\{\sum_j d_j \log \sum_{k=1}^{K} \theta_k \beta_{kj}; \quad -\frac{1}{2}(\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\log \tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}) - \sum_{k=1}^{K} \log \theta_k\}$
    $F := \frac{1}{\ell} \sum_{h=1}^{\ell} f_h$
    $i' := \arg \max_i \nabla F(\boldsymbol{\theta}_\ell)_i$; (maximal partial gradient)
    $\alpha := 2/(\ell + 2)$;
    $\boldsymbol{\theta}_{\ell+1} := \alpha \boldsymbol{e}_{i'} + (1 - \alpha)\boldsymbol{\theta}_\ell$.

end for

Return $\boldsymbol{\theta}_*$ with largest f amongst $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ...$

5.2 Why may Online Frank-Wolfe help?

We now discuss why OFW can do inference well in CTM even though inference is generally non-concave. In our observation, good performance of OFW originates mainly from (1) the probable concavity of the inference problem for which many instances in practice are concave, and from (2) the stochastic nature that allows OFW to get out of local optima to get closer to global ones.

Note that Step 1 of the learning algorithm has to do inference many times, each for a specific document. Hence, we have a family of inference instances. The analysis in Section 4 reveals that under some conditions, inferring topic mixtures in CTM is in fact concave. In

other words, there may be many concave instances in Step 1. For them, OFW is guaranteed to converge to the optimal solutions (Hazan and Kale, 2012).

For non-concave instances, inference is more difficult as there might be many local optima. Nonetheless, OFW is able find good approximate solutions due to at least two reasons. First, OFW is able to get out of local optima to reach closer to the global ones owing to its stochastic nature in selecting directions. Such an ability is intriguing that is missing in traditional deterministic algorithms for non-concave optimization. Second, due to the greedy nature, OFW is able to get close to local optima.

5.3 Experiments

This section is dedicated to answering the following three questions. *(a) How fast does fCTM do? (b) How good are the models learned by fCTM?* By answering these questions, we will see more clearly some benefits of studying probable convexity and the use of SGDs. *(c) How well and how fast does OFW resolve the inference problem in practice?* This question arises naturally as OFW (Hazan and Kale, 2012) was originally designed for concave problems while inference in CTM is non-concave in the worse case. Answer to this question also supports our highlight that SGDs might be a practical choice for non-concave optimization.

Four benchmark datasets were used in our investigation: KOS with 3430 documents, NIPS with 1500 documents, Enron with 39861 documents, and Grolier with 29762 documents.[2] For each dataset, we used 80% for learning models, and the remaining part was used to check the quality and efficiency of OFW.

5.3.1 How fast does fCTM perform?

To answer the first two questions and to see advantages of our algorithm (fCTM), we took the variational Bayesian method (denoted as CTM) by Blei and Lafferty (2007) into comparison. We used the same convergence criteria for fCTM and CTM: relative improvement of objective functions is less than 10^{-6} for inference of each document, and 10^{-3} for learning; at most 100 iterations are allowed to do inference. We used default settings for some other parameters of CTM. To avoid doing cross-validation for selecting the best value of α in fCTM, we used $\alpha = 1$ as the default setting.

Figure 1 records some statistics from learning and inference. We observed that fCTM learns significantly faster than CTM. Similar behavior holds when doing inference for each document. In our observations, fCTM often learns 60-170 times faster than CTM. Speedy learning of fCTM can be explained by the fact that Step 1 is done efficiently by OFW which has a linear convergence rate, provided that the inference problem is concave. In the cases of non-concave problems, OFW is still able to find efficiently approximate solutions. We observe that OFW often works 50-170 times faster than the variational method. In contrast, CTM did slowly because many auxiliary parameters need to be optimized when doing inference for each document. Furthermore, the variational method is not guaranteed to converge quickly. Figure 1 shows that CTM often needs intensive time to do inference.

Convergence speed: The last two rows in Figure 1 show how fast CTM and fCTM can reach convergence. Both methods can reach convergence in a relatively few iterations. We

---

2. KOS, NIPS, and Enron were retrieved from `http://archive.ics.uci.edu/ml/datasets/`. Grolier was retrieved from `http://cs.nyu.edu/~roweis/data.html`
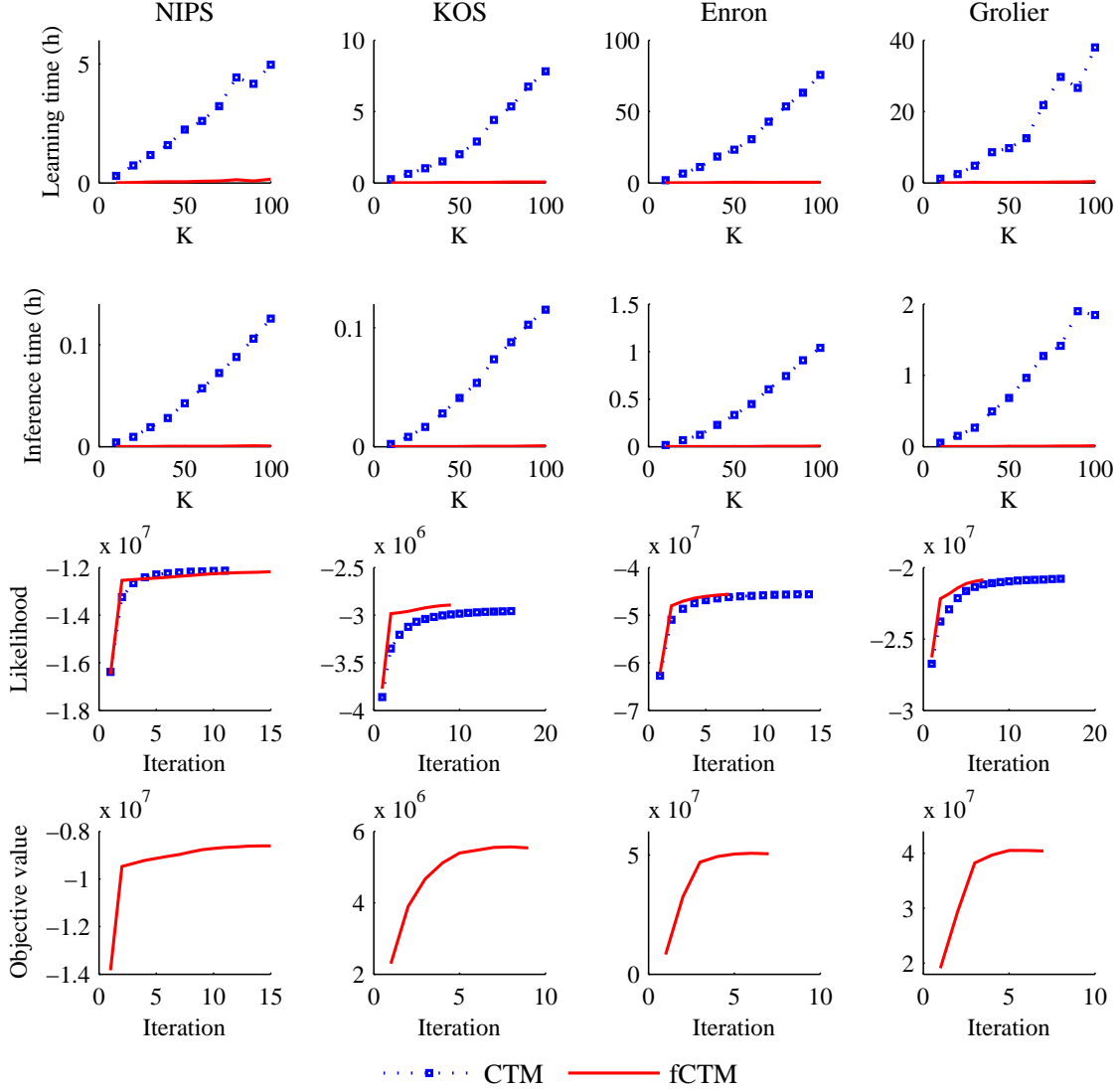
Figure 1: Performance of fCTM and CTM as the number K of topics increases. Lower is better for inference/learning time, whereas higher is better for likelihood. The last two rows show how fast fCTM can reach convergence for $K = 100$. We observe that fCTM often learns 60-170 times faster than CTM.
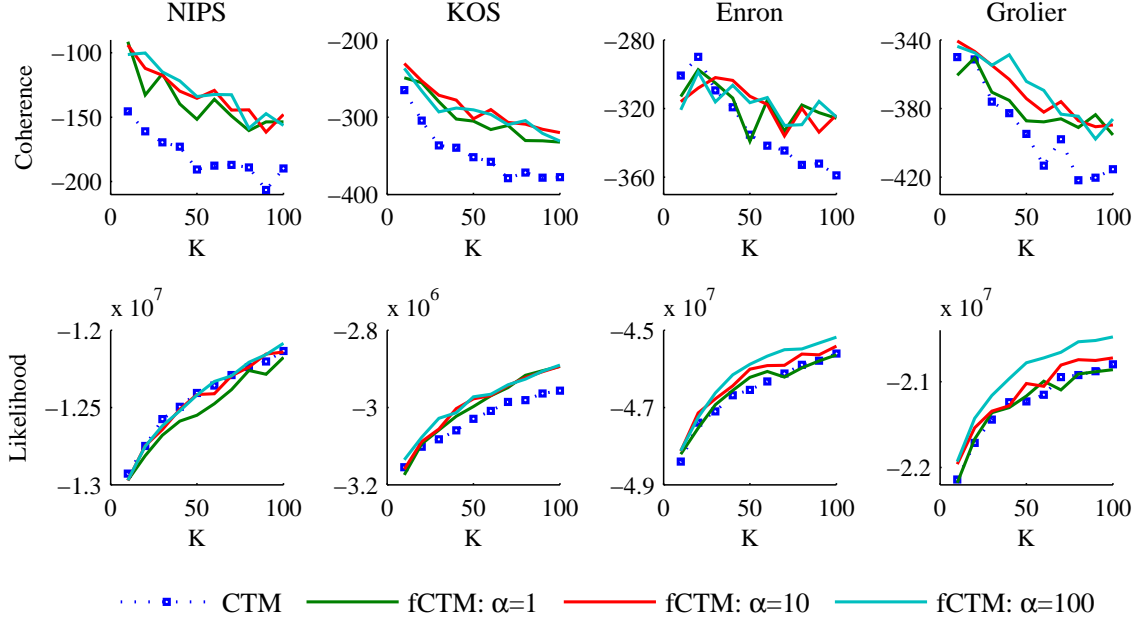
Figure 2: Quality of the models which were learned by fCTM (solid lines) and CTM (dashed lines). Higher is better.

observe that both methods rarely need 20 iterations to reach convergence; they both can reach stable after 10 iterations. Such a behavior would be beneficial when working in the cases of limited time.

5.3.2 HOW GOOD ARE THE MODELS LEARNED BY FCTM?

Likelihood and *coherence* are used to see the quality of models learned from data. Coherence is used to assess quality (goodness and interpretability) of individual topics. It has been observed to reflect well human assessment (Mimno et al., 2011).

To calculate the coherence of a topic k, we first choose the set $V^k = \{v_1^k, ..., v_t^k\}$ of the top t terms that have highest probabilities in that topic, and then compute

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

where $D(v)$ is the document frequency of term v, $D(u, v)$ is the number of documents that contain both terms u and v. In our experiments, we chose top $t = 20$ terms for investigation, and coherence of individual topics is averaged:

$$coherence = \frac{1}{K} \sum_{k=1}^{K} C(k, V^k).$$

Figure 2 shows the quality of the learned models. We observe that the two learning methods performed comparably in terms of likelihood. Note from Figure 1 that fCTM is
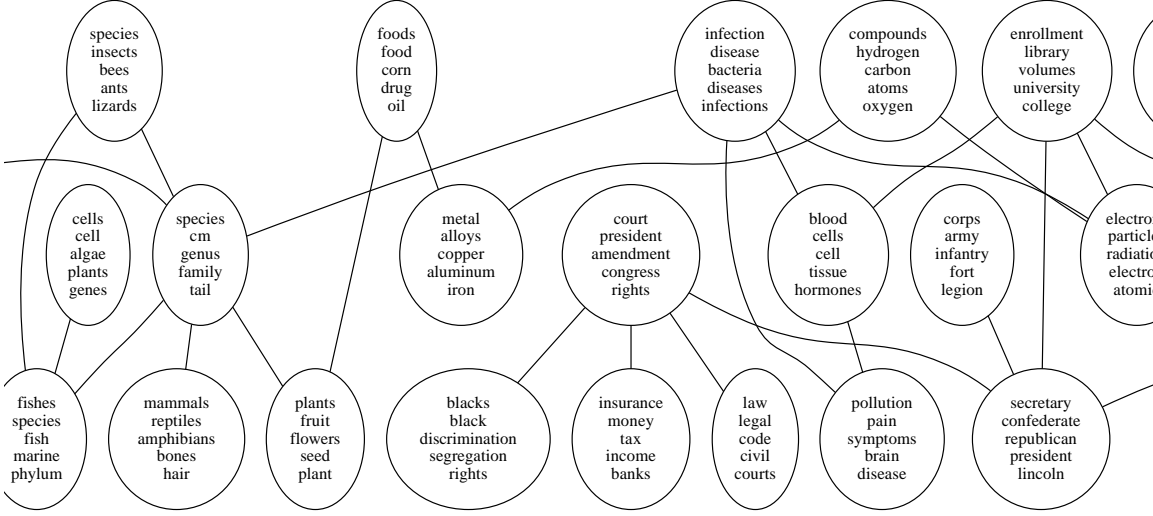
Figure 3: Illustration of the correlated topic model with 100 topics which was learned by fCTM from Grolier articles. An edge connecting two topics shows that if one topic appears in a document, the other likely appears as well. This visualization was drawn with Graphviz (Gansner and North, 2000)

.

able to reach comparable likelihood to CTM within few iterations, even though the main objective function of fCTM for learning is not likelihood. This behavior shows further the advantage of our algorithm.

In terms of coherence, the topic quality of CTM seems to be inferior to that of fCTM. Both methods often tend to learn less interpretable (but more specific) topics as the number K increases. CTM seems to degrade topic quality faster than fCTM as increasing K. We observe further that fCTM often learns significantly better topics than CTM in the cases of large K. When investigating the models learned by fCTM, we find that individual topics are very meaningful as depicted partially in Figure 3. Those observations demonstrate advantages of fCTM over CTM for practical applications, such as exploration or discovery of interactions of hidden topics/factors.

Models of hidden interactions: Figure 3 and 4 shows parts of the full model with 100 topics learned by fCTM from Grolier. Figure 3 shows positive correlations between topics, while Figure 4 shows negative correlations. It can be observed that the learned topics are interpretable and the discovered correlations are reasonable. Those further support that fCTM is able to learn qualitative models.

5.3.3 Quality and speed of OFW

We have seen in the last parts that OFW (an example of SGD algorithms) is really beneficial in helping fCTM to work efficiently. It seems to have more advantages than variational
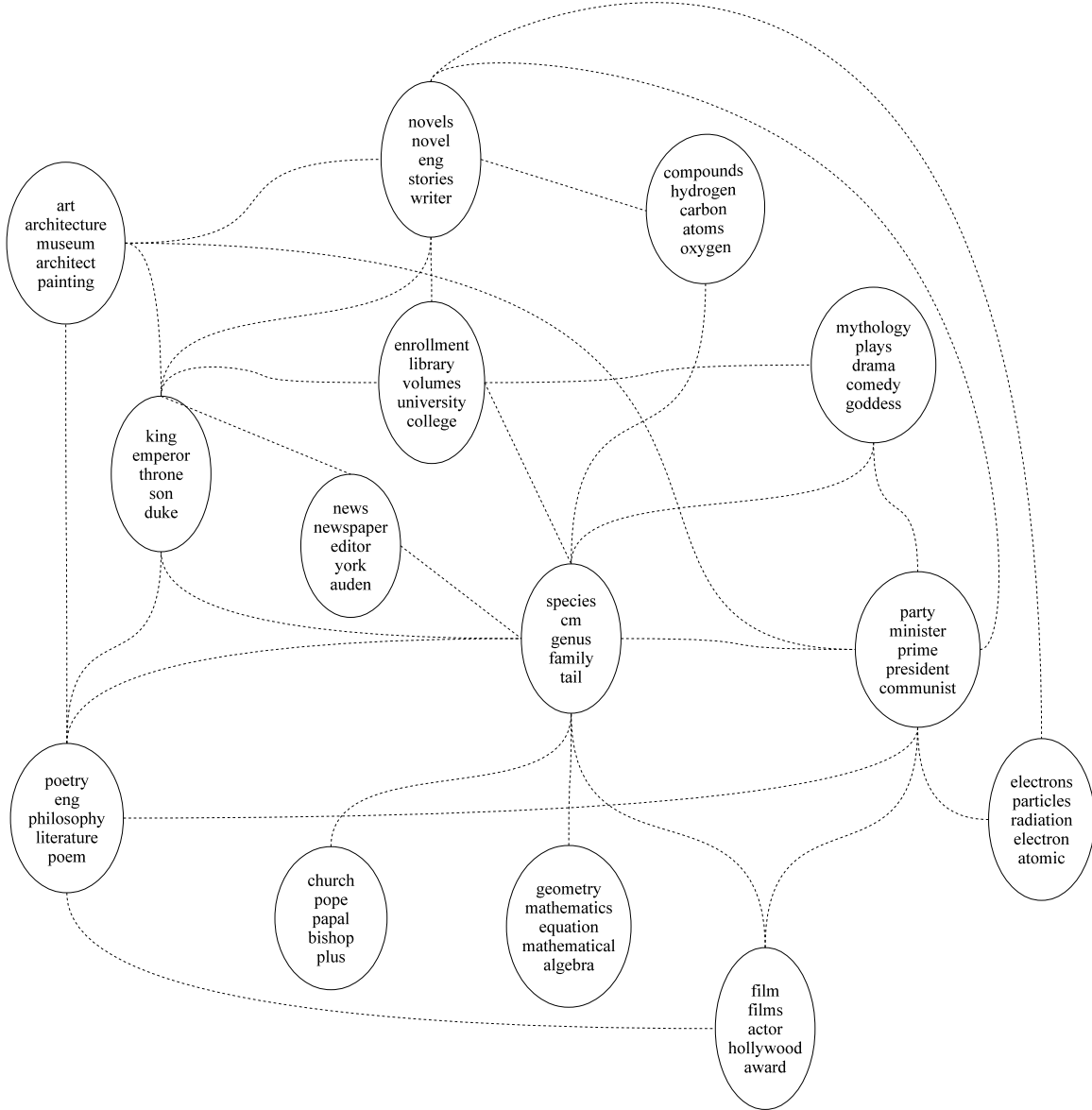
Figure 4: Illustration of the correlated topic model with 100 topics which was learned by fCTM from Grolier. An edge connecting two topics shows that the two topics *unlikely appear together* in a document.

Table 1: Statistics of OFW and SLSQP after solving 11197 non-concave problems. "#Fails to solve" shows the number of problems that a method found infeasible solutions. The last three rows show the number of problems that a method performs better than ($>$) or comparably with ($\approx$) or worse than ($<$) the other one. We observe that OFW often performs 150-2000 times faster than SLSQP. For most problems of interest, OFW found significantly better solutions than SLSQP.

Data		NIPS	KOS	Enron	Grolier
Total number of problems		150	343	3986	6718
Average time (seconds)	SLSQP	18.6729	2.2018	1.3871	1.5579
to solve a problem	*OFW*	*0.0161*	*0.0069*	*0.0056*	*0.0073*
Objective value (averaged)	SLSQP	-9708.9410	-169.3041	117.7672	-127.7626
	OFW	*-8860.8170*	*1464.7165*	*1753.9598*	*1572.2495*
#Fails to solve	SLSQP	37	172	1981	3316
	OFW	*0*	*0*	*0*	*0*
#OFW > SLSQP		129	343	3979	6700
#OFW ≈ SLSQP		21	0	7	16
#OFW < SLSQP		0	0	0	2

methods when being employed in CTM. Next, we are interested in performance of OFW as an algorithm for non-concave problems.

Problem (7) was used for investigation. The testing parts of the datasets were used to provide documents ($\boldsymbol{d}$) for (7). We used the models ($\Upsilon = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$) which have 100 topics and have been learned previously from the training data. Totally, we have 11197 instances of problem (7) for investigation.

For comparison, we took *Sequential Least Squares Programming* (SLSQP) as a standard method for non-convex optimization (Perez et al., 2012). Various methods have been proposed, but SLSQP seems to be one among the best solvers according to different tests (Perez et al., 2012). Therefore it was taken in comparison with OFW.[3] The same criterion was used to assess convergence for both methods: relative improvement of objective functions is no better than 10^{-6}, and the number of iterations is at most 100.

Table 1 shows some statistics from our experiments. It can be observed that SLSQP often needs intensive time to solve a problem, while OFW consumes substantially less time. We observe that OFW often works 150-2000 times faster than SLSQP. Slow performance of SLSQP mainly comes from the need to solve many intermediate quadratic programming problems, each of which often requires considerable time in our observations. On con-

---

3. The variational method by Blei and Lafferty (2007) was not considered for comparison. The reason comes from the difference of problems to be solved. Indeed, OFW tries to maximize $\Pr(\boldsymbol{\theta}, \boldsymbol{d})$ whereas the variational method tries to maximize a lower bound of the likelihood of document $\boldsymbol{d}$. Hence it is difficult to compare quality of the two methods. The last subsection has discussed inference time of the two methods.

trary, each iteration of OFW is very modest, which mostly requires computation of partial derivatives.

In terms of quality, we observe that OFW was able to find significantly better approximate solutions than SLSQP. When inspecting individual problems, we found that SLSQP failed to find feasible solutions for many problems, e.g., a large number of returned solutions were significantly out of domain ($\overline{\Delta}_K$). In contrast, OFW always manages to find feasible solutions. Among 11197 problems, OFW performed significantly worse than SLSQP for only 2. Those observations demonstrate that OFW has many advantages over (deterministic) SLSQP. Further, it is able to find good approximation solutions for non-concave problems with a modest requirement of computation.

6. Conclusion and discussion

We have introduced the concept of probable convexity to analyze real functions or families of functions. It is the way to see how probable a real function is convex. In particular, it can reveals how many members of a family of functions are convex. When a family contains most convex members, we could deal with the family efficiently in practice. Hence probable convexity provides a feasible way to deal with non-convexity of real problems such as posterior estimation in probabilistic graphical models.

When analysing probable convexity of the problem of estimating topic mixtures in CTM (Blei and Lafferty, 2007), we found that this problem is concave under certain conditions. The same results were obtained for many nonconjugate models. These results suggest that posterior inference of topic mixtures in those models might be done efficiently in practice, which seems to contradict with the belief of intractability in the literature. Benefiting from those theoretical results, we proposed a novel algorithm for learning CTM which can work 60-170 times faster than the variational method by Blei and Lafferty (2007), while keeping or making better the quality of the learned models. We believe that by using the same methodology as ours, learning for many existing nonconjugate models can be significantly accelerated. An implementation of our algorithm is freely available at `http://is.hust.edu.vn/~khoattq/codes/fCTM/`

There is a unusual employment of the Online Frank-Wolfe algorithm (OFW) (Hazan and Kale, 2012) to solve nonconvex problems (inference of topic mixtures in CTM). OFW is a specific instance of stochastic gradient descent algorithms (SGDs) for solving convex problems. By a careful employment, OFW behaves well in solving the inference problem which is nonconcave in the worse case. It helps us to design an efficient and effective algorithm for learning CTM. Such a successful use of OFW suggests that SGDs might be a practical choice to deal with nonconvex problems. In our experiments, OFW found significantly better solutions whereas performed 150-2000 times faster than SLSQP (the standard algorithm for nonconvex optimization). This further supports our highlight about SGDs. We hope that this highlight would open various rooms for future studies on connection of SGDs with nonconvex optimization.

References

Karim M. Abadir and Jan R. Magnus. *Matrix Algebra*. Cambridge University Press, 2005.

Amr Ahmed and Eric Xing. On tight approximate inference of the logistic-normal topic admixture model. In *AISTATS*, volume 2 of *Journal of Machine Learning Research: W&CP*, pages 19–26, 2007.

J Aitchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

David M. Blei and John Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Trans. Algorithms*, 6:63:1–63:30, 2010. ISSN 1549-6325. doi: http://doi.acm.org/ 10.1145/1824777.1824783. URL http://doi.acm.org/10.1145/1824777.1824783.

Shay B Cohen and Noah A Smith. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82. ACL, 2009.

Shay B Cohen and Noah A Smith. Covariance in unsupervised learning of probabilistic grammars. *The Journal of Machine Learning Research*, 11:3017–3051, 2010.

Emden R. Gansner and Stephen C. North. An open graph visualization system and its applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, 30(11):1203–1233, 2000.

T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th Annual International Conference on Machine Learning (ICML)*, 2012.

Gengxin Miao, Ziyu Guan, Louise E. Moser, Xifeng Yan, Shu Tao, Nikos Anerousis, and Ji-meng Sun. Latent association analysis of document pairs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1415–1423, New York, NY, USA, 2012. ACM. doi: 10.1145/2339530.2339752. URL http://doi.acm.org/10.1145/2339530.2339752.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference*

on Empirical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics, 2011.

Ruben E. Perez, Peter W. Jansen, and Joaquim R. R. A. Martins. pyOpt: A python-based object-oriented framework for nonlinear constrained optimization. *Structures and Multidisciplinary Optimization*, 45(1):101–118, 2012. doi: 10.1007/s00158-011-0666-3.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

D. Putthividhya, H. T. Attias, and S. Nagarajan. Independent factor topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

D. Putthividhya, H.T. Attias, and S.S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408 –3415, 2010. doi: 10.1109/CVPR.2010.5540000.

Konstantin Salomatin, Yiming Yang, and Abhimanyu Lad. Multi-field correlated topic modeling. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 628–637. SIAM, 2009.

David Sontag and Daniel M. Roy. Complexity of inference in latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Khoat Than and Tu Bao Ho. Fully sparse topic models. In Peter Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7523 of *Lecture Notes in Computer Science*, pages 490–505. Springer, 2012.

Joel Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Avi Wigderson and David Xiao. Derandomizing the ahlswede-winter matrix-valued chernoff bound using pessimistic estimators, and applications. *Theory of Computing*, 4(1):53–76, 2008.

Zhong P. Yang and Xiao X. Feng. A note on the trace inequality for products of hermitian matrix power. *Journal of Ineuqualities in Pure and Applied Mathematics*, 3(5):78:1–78:12, 2002.