2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

DIFFERENCE-OF-CONVEX OPTIMIZATION FOR VARIATIONAL KL-CORRECTED INFERENCE IN DIRICHLET PROCESS MIXTURES

Rasmus Bonnevie, Morten Mørup, Mikkel N. Schmidt

Technical University of Denmark
Department of Applied Mathematics and Computer Science

ABSTRACT

Variational methods for approximate inference in Bayesian models optimise a lower bound on the marginal likelihood, but the optimization problem often suffers from being non-convex and high-dimensional. This can be alleviated by working in a collapsed domain where a part of the parameter space is marginalized. We consider the KL-corrected collapsed variational bound and apply it to Dirichlet process mixture models, allowing us to reduce the optimization space considerably. We find that the variational bound exhibits consistent and exploitable structure, allowing the application of difference-of-convex optimization algorithms. We show how this yields an interpretable fixed-point update algorithm in the collapsed setting for the Dirichlet process mixture model. We connect this update formula to classical coordinate ascent updates, illustrating that the proposed improvement surprisingly reduces to the traditional scheme.

Index Terms— difference-of-convex optimization, variational inference, collapsed methods, bayesian nonparametrics

1. INTRODUCTION

Although variational inference has been around for a while [1], there has been a surge in interest lately, moving variational inference beyond the traditional mean-field approximation and coordinate-ascent optimization. Recent advances include algorithms for non-conjugate black box inference [2], stochastic optimization in the large data setting [3], and universally applicable probabilistic programming software [4], making inference tractable for complex models such as Bayesian neural networks [5].

Despite these advances, the variational approach hinges on solving a potentially massive, non-convex, and high-dimensional optimization problem. Reducing the parameter space by analytically marginalizing parts of the variational approximation can lead to a more well-behaved objective function, faster convergence, and better solutions [6]. To this end, we adopt the KL-corrected (KLC) bound as our variational objective. It was originally invented for Gaussian processes alone [7], but was later extended to a larger class of conjugate exponential models [8] where it was demonstrated to reduce the optimization space in a principled manner without affecting the set of solutions. Furthermore, it has already been shown to lead to more efficient optimization [8].

Our primary contribution is the realization that the KLC bound has consistent structure when applied to a Dirichlet process mixture, as it decomposes nicely into convex and concave terms. This leads us to consider difference-of-convex (DC) optimization as exemplified in the convex-concave procedure [9] and its generalization to non-differentiable objectives, the aptly named Difference-of-Convex Algorithm (DCA) [10]. We show that this leads to a nice fixed-point mapping which can be expressed as the softmax of a gradient related to the joint distribution.

While superficially different, and derived by a different route, this fixed-point formula turns out to reduce to the classical mean-field update. We investigate under which conditions this holds and find that it is symptomatic of models with exponential family conditionals. We consider the perspectives of this alternate derivation, including how results about convergence can potentially be carried over.

2. THE KL-CORRECTED VARIATIONAL LOWER BOUND

Consider the general Bayesian problem of inferring a distribution over latent variables $\boldsymbol{Z}$ and internal (nuisance) parameters $\boldsymbol{U}$ given observations of a random variable $\boldsymbol{X}$. We can compute the posterior $p(\boldsymbol{Z}, \boldsymbol{U}|\boldsymbol{X})$ up to a constant, but the normalization constant is typically intractable. Variational inference gets around this issue by defining a family of approximations $q(\boldsymbol{Z}, \boldsymbol{U})$ and then minimizing the KL divergence $\mathrm{KL}(q\|p)$. The KL divergence is similarly intractable, but it shares its critical points with the standard variational lower bound:

$$\mathcal{L}_{MF} = \mathbb{E}_q[\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U})] - \mathbb{E}_q[\ln q(\boldsymbol{Z}, \boldsymbol{U})]\,. \quad (1)$$

Minimizing this non-convex objective with respect to the parameters of q leads to a locally optimal approximation of $p(\boldsymbol{Z}, \boldsymbol{U}|\boldsymbol{X})$. Equation (1) is referred to as a lower bound as it lower bounds the log-evidence $\ln p(\boldsymbol{X})$.

978-1-5090-6341-3/17/$31.00

Suppose we now try to marginalize $\boldsymbol{U}$ prior to doing inference, then the resulting bound has the form

$$\mathcal{L}_C = \mathbb{E}_q\left[\ln \int p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U})\, \mathrm{d}\boldsymbol{U}\right] - \mathbb{E}_q[\ln q(\boldsymbol{Z})]\,, \quad (2)$$

which unfortunately requires the computation of the expectation of a log-integral. Even if the integral is tractable, the expectation over q often will not be. In the particular case of conjugate exponential family models the integral leads to a compound distribution outside of the exponential family, which means that we lose many of the tractability benefits of working with exponential family models.

This brings us to the KL-corrected bound. The derivation of the KL-corrected lower bound is a form of pseudo-marginalization which reduces the parameter space and leaves a more well-behaved (and still tractable) objective function, but where the inference is still effectively over the original unmarginalized model. There are several ways to derive it, and we will follow Hensman et al. by deriving it by way of an auxiliary bound [8].

The auxiliary bound is derived as a standard lower bound, but for the model conditioned on $\boldsymbol{U}$, instead of on the full joint distribution.

$$\mathcal{L}_1(\boldsymbol{U}) = \mathbb{E}_{q(\boldsymbol{Z})}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{U}) - \ln q(\boldsymbol{Z})]\,. \quad (3)$$

Note that the bound is a function of $\boldsymbol{U}$. The conditional bound can be transformed into the KL-corrected bound as follows

$$\mathcal{L}_{KL} = \ln \mathbb{E}_{p(\boldsymbol{U})}[\exp(\mathcal{L}_1(\boldsymbol{U}))]\,. \quad (4)$$

Since $\mathcal{L}_1(\boldsymbol{U})$ is a bound on $\ln p(\boldsymbol{X}|\boldsymbol{U})$, the operations above result in $\mathcal{L}_{KL}$ being a bound on the marginal likelihood $\ln p(\boldsymbol{X})$ as desired. The KLC bound is related to the CVB0 approximation [11] as detailed in the original article [8].

2.1. The KLC Bound for the Dirichlet Process Mixture

As an example, we will consider a particular conjugate exponential family model where the KL-corrected bound is computationally advantageous — namely a Dirichlet process mixture model. KLC bounds for finite mixture models have already been covered [8, supplementary], but we will need the KLC bound later so we provide the derivation here for the non-parametrically extended mixture. We will leave the component distribution arbitrary, under the constraint that it is an exponential family distribution with a density of the form

$$p(\boldsymbol{x}_i|\boldsymbol{\eta}_k) = h(\boldsymbol{x}_i)\exp\big(\boldsymbol{\eta}_k^\top T(\boldsymbol{x}_i) - A(\boldsymbol{\eta}_k)\big) \quad (5)$$

We further model each parameter vector $\boldsymbol{\eta}_k$ as being drawn from a common conjugate prior $\boldsymbol{\eta}_k \sim p(\boldsymbol{\eta}|\boldsymbol{\gamma}, \nu)$. We can combine the above into a mixture model using latent indicators $\boldsymbol{Z}$

$$p(\boldsymbol{X}, \{\boldsymbol{\eta}_k\}_{k=1}^\infty|\boldsymbol{Z}) = \prod_{k=1}^\infty \left[\prod_{i=1}^N p(\boldsymbol{x}_i|\boldsymbol{\eta}_k)^{z_{ik}}\right] p(\boldsymbol{\eta}_k|\boldsymbol{\gamma}, \nu) \quad (6)$$

The final component needed is the prior on $\boldsymbol{Z}$. We will employ a stick-breaking representation of the Dirichlet process, but one could just as well use a Dirichlet-multinomial pair for a finite mixture. The stickbreaking distribution takes the form

$$\beta_k \sim \mathrm{Beta}(1, \alpha), \quad z_{ik}|\beta_1, \ldots, \beta_k \sim \mathrm{Cat}(w_k) \quad (7)$$

with dependency through $w_k = \beta_k \prod_{\ell<k}(1-\beta_\ell)$. While this prior has an unbounded number of variables, we will control for this later using the variational approximation so that the bounds only ever have a finite number of terms. Writing out the prior density gives us

$$p(\boldsymbol{Z}|\boldsymbol{\beta}) = \prod_{i=1}^N \prod_{k=1}^\infty \left[\beta_k \prod_{\ell<k}(1-\beta_\ell)\right]^{z_{ik}} = \prod_{k=1}^\infty \beta_k^{m_k}(1-\beta_k)^{m_{k+1}^\infty}, \quad (8)$$

where $m_k = \sum_{i=1}^N z_{ik}$ and $m_{k+1}^\infty = \sum_{\ell=k+1}^\infty m_\ell$.

Collecting $\boldsymbol{U} = (\{\boldsymbol{\eta}_k\}_{k=1}^\infty, \boldsymbol{\beta})$, we can compute the $\mathcal{L}_1(\boldsymbol{U})$ conditional bound, under an exponential family variational approximation $q(\boldsymbol{Z}|\boldsymbol{\mu})$ parametrized by mean parameters $\boldsymbol{\mu}$, resulting in

$$\mathcal{L}_1(\boldsymbol{U}) = C + \sum_{k=1}^\infty \big[\boldsymbol{\eta}_k^\top \bar{T}_k - \bar{m}_k A(\boldsymbol{\eta}_k)\big] + \sum_{k=1}^\infty \big[\bar{m}_k \ln(\beta_k) + \bar{m}_{k+1}^\infty \ln(1-\beta_k)\big] + \mathcal{H}_q(\boldsymbol{\mu}), \quad (9)$$

where $C = \sum_{i=1}^N \ln h(\boldsymbol{x}_i)$, $\bar{T}_k = \sum_{i=1}^N \mathbb{E}_q[z_{ik}]\, T(\boldsymbol{x}_i)$, $\bar{m}_k = \sum_{i=1}^N \mathbb{E}_q[z_{ik}]$, and $\bar{m}_{k+1}^\infty = \sum_{\ell=k+1}^\infty \bar{m}_\ell$.

To compute the KL-corrected bound we will split up the $\mathcal{L}_1$ bound into terms depending on $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$. Taking the exponential as in equation (4) gives us expressions with the same functional forms as the original distributions, allowing us to integrate over the appropriate conjugate priors, yielding factors

$$\prod_{k=1}^\infty \frac{e^{A_0(\boldsymbol{\gamma}+\bar{T}_k, \nu+\bar{m}_k)}}{e^{A_0(\boldsymbol{\gamma},\nu)}}, \quad \prod_{k=1}^\infty \frac{B(1+\bar{m}_k, \alpha+\bar{m}_{k+1}^\infty)}{B(1,\alpha)}. \quad (10)$$

for the likelihood and latent distributions, respectively. Here, A_0 is the log-normalizer of the conjugate prior to $\boldsymbol{\eta}$ and B is the beta function, i.e. the normalizer of the Beta distribution..

The KL-corrected bound now follows naturally from the definition.

$$\mathcal{L}_{KL} = \sum_{k=1}^\infty \big[A_0(\boldsymbol{\gamma}+\bar{T}_k, \nu+\bar{m}_k) + \ln B(1+\bar{m}_k, \alpha+\bar{m}_{k+1}^\infty)\big] + \mathcal{H}_q(\boldsymbol{Z}) + \text{const.} \quad (11)$$

2.2. Difference-of-Convex Structure of the KLC Bound

Our key observation is that both A_0 and $\ln B$ are the log-normalizers of exponential family models and are thus known to be convex [12]. Since $\bar{T}_k$ and $\bar{m}_k$ are linear functions in $\mu_{ik} \equiv \mathbb{E}_q[z_{ik}]$, we know that their composition with a convex function results in something that is also convex in μ_{ik} [13]. Since the sum likewise preserves convexity, the whole sum is convex.

Furthermore, if $q(\boldsymbol{Z})$ is an exponential family with mean parametrization $\boldsymbol{\mu}$ and log-normalizer A_q, then it can also be shown that [12, theorem 3.4]

$$-A_q^*(\boldsymbol{\mu}) = \mathcal{H}_q(\boldsymbol{\mu}), \tag{12}$$

where A_q^* denotes the convex conjugate of the log-normalizer. Since the convex conjugate is always convex, we can conclude that the entropy is concave for an exponential family [12].

To summarize, the bound is made up of a convex and a concave part; additional structure we should do our best to exploit. To stay true to the optimization literature, we will consider minimization of $-\mathcal{L}_{KL}$ from here on out, resulting in the following (flipped) decomposition

$$-\mathcal{L}_{KL} = f_{vex} + f_{cave} - C, \quad f_{vex} = -\mathcal{H}_q(\boldsymbol{\mu}), \tag{13}$$

$$f_{cave} = -\sum_{k=1}^{\infty} \left[A_0(\boldsymbol{\gamma} + \bar{T}_k, \nu + \bar{m}_k) + \right. \tag{14}$$

$$\left. \ln B(1 + \bar{m}_k, \alpha + \bar{m}_{k+1}^{\infty}) \right]. \tag{15}$$

3. CONVEX-CONCAVE PROCEDURE

Optimization problems with a mix of convex and concave terms are denoted as difference-of-convex problems (DC). Technically, any non-convex smooth problem is a DC problem as functions can be decomposed into regions of positive and negative curvature, but the decomposition is not always obvious [9, 10].

The **c**onvex-**c**on**c**ave **p**rocedure (CCCP) is a straightforward algorithm for DC problems [9]. The core idea is that a stationary point for a difference function occurs when the gradients of the two terms match, i.e.

$$0 = \nabla \left(f_{vex} + f_{cave} \right) \Leftrightarrow \nabla f_{vex} = -\nabla f_{cave}. \tag{16}$$

The CCCP algorithm simply turns this premise into an implicit fixed-point scheme

$$\nabla f_{vex}(\boldsymbol{\mu}_{t+1}) = -\nabla f_{cave}(\boldsymbol{\mu}_t), \tag{17}$$

so $\boldsymbol{\mu}_{t+1}$ is picked so that the convex gradient matches the negative concave gradient at time t. While this might look arbitrary, this in fact elegantly exploits the features of the function's convex-concave nature, ensuring a monotonously decreasing sequence [9].

An equivalent, but slightly more approachable, interpretation of CCCP is as a sequential optimization problem, where

$$\boldsymbol{\mu}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\mu}} f_{vex}(\boldsymbol{\mu}) + \tilde{f}_{cave}^{(\boldsymbol{\mu}_t)}(\boldsymbol{\mu}). \tag{18}$$

where we have linearized the concave part around $\boldsymbol{\mu}_t$ as $\tilde{f}_{cave}^{(\boldsymbol{\mu}_t)}(\boldsymbol{\mu}) = \big(f_{cave}(\boldsymbol{\mu}_t) + (\boldsymbol{\mu} - \boldsymbol{\mu}_t)^\top \nabla f_{cave}(\boldsymbol{\mu}_t)\big)$. Since the concave part is linearized it becomes trivially convex (in addition to concave), and the complete objective is then unequivocally a convex function, leading to a simplified problem where the full brunt of convex optimization can be brought into play. Since the linearization of a concave function upper bounds the concave function itself this provides an illustration of why the sequence is monotonously decreasing (see figure 1). for more details, see [14, 9, 10].

3.1. Necessary Conditions on the Variational Distribution

So far, we have left the variational approximation $q(\boldsymbol{Z}|\boldsymbol{\mu})$ vague. With the above in place, we see that to apply CCCP to our bound, there are two key restrictions (and an additional facilitator).

Expectations Linear in the Parameters We are relying on the transparent relationship $\boldsymbol{\mu}_{ik} = \mathbb{E}[z_{ik}]$ between parameters and expected latent variables. This could be relaxed a bit — the expectation could be any linear function of the parameters. In fact, it could even be a convex or concave function of the parameters following the standard composition theorems, assuming some further conditions hold [13].

Concave Entropy The variational approximation needs to have concave entropy. The entropy function is concave for exponential family models [12], and entropy in general is concave in the space of distributions, but we have been unable to document that this holds for all distributions outside of the exponential family, as well as all possible parameterizations.

(Tractable Inverse Gradient Map) Ideally, we would also like to know the inverse entropy-gradient map. This turns out to be well-known for many exponential families, but is likely unavailable for many more interesting variational approximations. Fortunately, we will still be able to solve the sequential problem in equation (18) efficiently if q obeys the other conditions, so the inverse map is not strictly necessary.

We can find at least one simple variational approximation obeying the above conditions in the form of the widely used product of single-sample multinomials (i.e. categorical distributions).

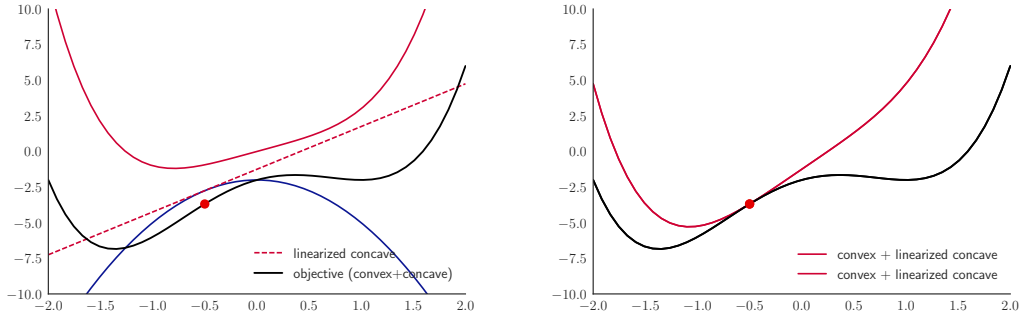

Fig. 1. The sequential interpretation minimizes the objective (black) by constructing an upper bound. The concave term is linearized to yield a convex upper bound (red; right).

3.2. Fixed-point Update for the KLC Bound

To formulate our main result, we rewrite equation (17) following [9],

$$\boldsymbol{\mu}_{t+1} = \left[\nabla f_{vex}\right]^{-1}\left(-\nabla f_{cave}(\boldsymbol{\mu}_t)\right). \tag{19}$$

Since the KLC bound consists of log-normalizers, we just need to be able to compute gradients of exponential family log-normalizers to compute the gradient of the bound. This makes it relevant to mention the following relationships between the (arbitrary exponential family) distribution's log-normalizer A, its natural parameters $\boldsymbol{\eta}$, and its dual parametrization in mean parameters $\boldsymbol{\mu}$ [12]

$$\left[\nabla A\right]^{-1}(\boldsymbol{\mu}) = \nabla A^*(\boldsymbol{\mu}) = \boldsymbol{\eta}, \tag{20}$$

$$\left[\nabla A^*\right]^{-1}(\boldsymbol{\eta}) = \nabla A(\boldsymbol{\eta}) = \boldsymbol{\mu}, \tag{21}$$

illustrating key symmetries found in exponential family models. Recall that $f_{vex} = -\mathcal{H}_q(\boldsymbol{\mu}) = A_q^*(\boldsymbol{\mu})$, so that $\left[\nabla f_{vex}\right]^{-1} = \nabla A_q$ by the above. Then CCCP yields

$$\boldsymbol{\mu}_{t+1} = \nabla A_q(-\nabla f_{cave}(\boldsymbol{\mu}_t)). \tag{22}$$

If we compare this to the second identity in (20), it appears that $-\nabla f_{cave}(\boldsymbol{\mu}_t)$ is in some sense representing a set of natural parameters. At the fixed point $\boldsymbol{\mu}^*$ of the update rule, it must in fact be the exact corresponding natural parameter $\boldsymbol{\eta}^*$, i.e. $-\nabla f_{cave}(\boldsymbol{\mu}^*) = \boldsymbol{\eta}^* = \nabla A_q^*(\boldsymbol{\mu}^*)$.

We can make the update rule a bit more explicit, but first we have to handle the normalization constraints $\sum_{k=1}^K \mu_{ik} = 1$. We add Lagrangian terms to the convex terms such that

$$f_{vex} = \sum_{i=1}^N \sum_{k=1}^K \mu_{ik} \ln \mu_{ik} + \sum_{i=1}^N \lambda_i \left(\sum_{k=1}^K \mu_{ik} - 1\right). \tag{23}$$

Taking the derivative yields

$$g_{ik} = \frac{\partial}{\partial \mu_{ik}} f_{vex} = \ln \mu_{ik} + 1 + \lambda_i \Leftrightarrow \mu_{ik} = \frac{\exp(g_{ik} - 1)}{\exp(\lambda_i)}. \tag{24}$$

Applying the constraint, we get that

$$\mu_{ik} = \frac{\exp(g_{ik})}{\sum_{k=1}^K \exp(g_{ik})}, \tag{25}$$

which is the softmax function $\mathcal{S}(\cdot)$, so we can write the actual CCCP update formula (equation (19)) as

$$\boldsymbol{\mu}_{t+1} = \mathcal{S}(-\nabla f_{cave}(\boldsymbol{\mu}_t)) = \mathcal{S}(\nabla_{\boldsymbol{\mu}} \ln p(\boldsymbol{X}, \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{Z}]))|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t}. \tag{26}$$

This is reminiscent of exponentiated gradient algorithms which show up when objectives are regularized with Kullback-Leibler divergences [15, 16]. Since the KL divergence includes an entropy term it makes sense that they appear similar.

4. CONNECTING THE KLC UPDATES WITH MEAN-FIELD COORDINATE ASCENT

The original solution to the mean-field variational inference problem in the uncollapsed setting was to apply coordinate ascent, updating each distribution in turn. This procedure was often tractable for conjugate exponential family models, if sometimes convoluted.

In general, by taking the derivative of the lower bound with respect to the parameters controlling the distribution over model variable $\theta_i \in \{\theta_j\}_{j=1}^N$ and setting the derivative to zero, we can find that the optimal variational approximation is[17]

$$q(\theta_i) \propto \exp(\mathbb{E}_q[\ln p(\theta_i|\mathcal{D}, \{\theta_j\}_{i\neq j}) \mid \theta_i]), \tag{27}$$

where $\mathcal{D}$ is the set of observed variables. Usually, mean field assumptions are exploited to ensure that the expectations are tractable, but if the expectations are computable without that assumption then the parameters can be updated in blocks.

For our mixture, the expectation resolves to

$$\mathbb{E}_q\big[\ln p(z_i|\boldsymbol{X}, \boldsymbol{Z}_{\setminus i}, \boldsymbol{U}) \mid \boldsymbol{Z}\big] = \tag{28}$$

$$\sum_{k=1}^{K} z_{ik} \left(T(\boldsymbol{x}_i)^\top \mathbb{E}[\boldsymbol{\eta}_k] - \mathbb{E}[A(\boldsymbol{\eta}_k)] + \mathbb{E}[\ln w_k(\boldsymbol{\beta})]\right) \tag{29}$$

where we will denote the term in the parenthesis by $\ln \tilde{\pi}_{ik}$, which is understood to be the log of the unnormalized probability parametrizing the multinomial variational approximation over z_i. If we take the softmax of $\ln \tilde{\pi}_{ik}$ we recover the distribution itself. Let us compare this to $-\nabla f_{cave}$. We will consider its terms individually, starting with the log-normalizer terms involving A_0. Taking the gradient, we get

$$\frac{\partial}{\partial \mu_{ik}} A_0(\boldsymbol{\gamma} + \bar{T}_k, \nu + \bar{m}_k) = \tag{30}$$

$$\nabla_{\boldsymbol{\gamma},\nu} A_0^\top \begin{pmatrix} T(\boldsymbol{x}_i) \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\boldsymbol{\eta}_k] \\ -\mathbb{E}[A(\boldsymbol{\eta}_k)] \end{pmatrix}^\top \begin{pmatrix} T(\boldsymbol{x}_i) \\ 1 \end{pmatrix}. \tag{31}$$

These terms exactly match the ones found in the coordinate ascent update.

Following a similar process, if we take the derivatives of the log-beta terms — being log-normalizers of beta distributions — in f_{cave}, we recover the latent terms in the ascent updates

$$\frac{\partial}{\partial \mu_{i\ell}} \ln B(1 + \bar{m}_k, \alpha + \bar{m}_{k+1}^{\infty}) = \begin{pmatrix} \mathbb{E}[\ln \beta_k] \\ \mathbb{E}[\ln(1-\beta_k)] \end{pmatrix}^\top \begin{pmatrix} 1[\ell = k] \\ 1[\ell < k] \end{pmatrix}, \tag{32}$$

as the latent terms in the ascent formula can be expanded as

$$\mathbb{E}[\ln w_k(\boldsymbol{\beta})] = \mathbb{E}[\ln \beta_k] + \sum_{\ell<k} \mathbb{E}[\ln(1-\beta_k)]. \tag{33}$$

Thus we can conclude that the CCCP strategy exactly matches classical coordinate ascent. This conclusion hinges on the expectations being over the same variational distribution $q(\boldsymbol{U})$, but the KL-corrected bound implicitly always uses the optimal approximation $q^*(\boldsymbol{U})$ so if we take coordinate ascent steps to maximize $q(\boldsymbol{U})$ before updating $\boldsymbol{\mu}$, then the expectations will always match.

To investigate this further, recall that in the uncollapsed setting we need to find both a variational approximation $q(\boldsymbol{U})$ over the component parameters, as well as a distribution over the clusters parameterized by $\boldsymbol{\mu}$. To every state $\boldsymbol{\mu}_t$, there is an optimal setting of the variational approximation $q(\boldsymbol{U})$; we use $\Lambda(\boldsymbol{\mu}_t)$ to denote the implicit map that maps $\boldsymbol{\mu}_t$ to the optimal $q^*(\boldsymbol{U})$. Let us use that the entropy $(-f_{vex})$ is a term $\mathcal{L}_{MF}$ and $\mathcal{L}_{KL}$ have in common and define a decomposition $-\mathcal{L}_{MF}(\boldsymbol{\mu}, \Lambda(\boldsymbol{\mu}_t)) = -\mathcal{E}(\boldsymbol{\mu}, \Lambda(\boldsymbol{\mu}_t)) + f_{vex}(\boldsymbol{\mu})$ where $\mathcal{E}$ is the average energy — the first argument $\boldsymbol{\mu}$ is identified with a distribution over $\boldsymbol{Z}$, while the second argument is the distribution over $\boldsymbol{U}$ which we set to the optimal value with respect to a previous iterate $\boldsymbol{\mu}_t$, using the implicit map $\Lambda(\cdot)$. If $\boldsymbol{\mu}_*$ is the optimum of the bound, the first-order optimality condition for $\mathcal{L}_{MF}$ with respect to $\boldsymbol{\mu}$ states that

$$\nabla \mathcal{E}(\boldsymbol{\mu}_*, \Lambda(\boldsymbol{\mu}_t)) = \nabla f_{vex}(\boldsymbol{\mu}_*) \tag{34}$$

Now recall that CCCP can be interpreted as a sequential convex problem (equation (18)) with a linearized concave component $\tilde{f}_{cave}(\boldsymbol{\mu})$. We then have the exact same optimality condition, but at a potentially different point: $-\nabla \tilde{f}_{cave}(\tilde{\boldsymbol{\mu}}_*) = \nabla f_{vex}(\tilde{\boldsymbol{\mu}}_*)$. Furthermore, recall that the two bounds have matching values and gradients at $\boldsymbol{\mu}_t$ by construction, i.e. $\nabla \mathcal{L}_{KL}(\boldsymbol{\mu}_t) = \nabla \mathcal{L}_{MF}(\boldsymbol{\mu}_t)$, which means the energy must match the linearized concave component

$$-\nabla \tilde{f}_{cave}(\tilde{\boldsymbol{\mu}}_*) = \mathcal{E}(\boldsymbol{\mu}_t, \Lambda(\boldsymbol{\mu}_t))$$

since $\tilde{f}_{cave}$ is linear, its gradient is constant, so if $\tilde{\boldsymbol{\mu}}_* = \boldsymbol{\mu}_*$, we have the peculiar property that $\nabla \mathcal{E}(\boldsymbol{\mu}_t, \Lambda(\boldsymbol{\mu}_t)) = \nabla \mathcal{E}(\boldsymbol{\mu}_*, \Lambda(\boldsymbol{\mu}_t))$, i.e. the gradient of the energy is constant as well. This hints at "hidden linearity" in the average energy.

A partial explanation comes from considering the case where the distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{U})$ is in an exponential family together with its prior $p(\boldsymbol{Z}|\nu)$. Following Hoffman et al. [3], the *natural* gradient $\tilde{\nabla}$ of the lower bound for an exponential family with parameters $\boldsymbol{\eta}$ is

$$\tilde{\nabla}_{\boldsymbol{\eta}} \mathcal{L}_{MF}(\boldsymbol{\eta}) = \boldsymbol{\eta} - \mathbb{E}_{q(\nu)}[\nu] = \nabla_{\boldsymbol{\mu}} \mathcal{L}_{MF}(\boldsymbol{\eta}(\boldsymbol{\mu})), \tag{35}$$

where the last equality can be proven using the chain rule [8]. Since the $\boldsymbol{\eta}$ term comes from the entropy, the natural gradient in the energy does indeed appear to be constant. So from this it's clear that mean parameter gradients of the average energy are constant when the conditionals are exponential family distributions.

We note that this identity between the two algorithms has its benefits and can provide new angles of attack for theoretical problems concerning variational inference. As an example, convergence for CCCP and other bound optimization algorithms was investigated by Salakhutdinov et al. [14]. Finally, we should mention that this is not the first connection found between the coordinate ascent updates and other optimization paradigms. Sato deduced that the coordinate ascent updates were similarly identical to natural gradient steps with stepsize 1 [18]. By transitivity our iteration formula is then also identical to a unit natural gradient step.

5. CONCLUSION

The main result of this paper is the demonstration that the KL-corrected bound for the Dirichlet process mixture inher-

its structure from the original variational problem and can be partitioned into convex and concave parts.

We argue that additional information available about an objective function should be exploited to the extent possible, and the difference-of-convex literature indicates that the above split can lead to improved non-convex optimization.

Applying the CCCP algorithm leads to a general analytical fixed-point update formula. The update formula is shown to match standard variational Bayes updates, and thus provides a new angle of attack on the variational problem, which can potentially be extended to models beyond the classical mixture model.

To truly surpass the existing inference schemes it appears that we need difference-of-convex algorithms that can take advantage of second-order derivatives. Unfortunately, to the best of our knowledge, the DC optimization literature has yet to find algorithms improving on CCCP/DCA. We hope that future research will either uncover new ways to exploit the difference-of-convex structure, or that the connections with DC optimization can provide a new fruitful avenue for the analysis of collapsed variational Bayes.

6. REFERENCES

[1] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1 Nov. 1999.

[2] Rajesh Ranganath, Sean Gerrish, and David M Blei, "Black box variational inference," 31 Dec. 2013.

[3] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *Journal of machine learning research: JMLR*, vol. 14, no. 1, pp. 1303–1347, May 2013.

[4] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei, "Automatic differentiation variational inference," 2 Mar. 2016.

[5] Diederik P Kingma and Max Welling, "Auto-Encoding variational bayes," 20 Dec. 2013.

[6] Yee W Teh, David Newman, and Max Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 19*, B Schölkopf, J C Platt, and T Hoffman, Eds., pp. 1353–1360. MIT Press, 2007.

[7] Nathaniel J King and Neil D Lawrence, "Fast variational inference for gaussian process models through KL-Correction," in *Machine Learning: ECML 2006*, Lecture Notes in Computer Science, pp. 270–281. Springer Berlin Heidelberg, 1 Jan. 2006.

[8] James Hensman, Magnus Rattray, and Neil D Lawrence, "Fast variational inference in the conjugate exponential family," in *Advances in Neural Information Processing Systems 25*, F Pereira, C J C Burges, L Bottou, and K Q Weinberger, Eds., pp. 2888–2896. Curran Associates, Inc., 2012.

[9] A L Yuille and Anand Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[10] Le Thi Hoai An and Pham Dinh Tao, "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Annals of Operations Research*, vol. 133, no. 1-4, pp. 23–46, 2005.

[11] Katsuhiko Ishiguro, Issei Sato, and Naonori Ueda, "Averaged collapsed variational bayes inference," *Journal of machine learning research: JMLR*, vol. 18, no. 1, pp. 1–29, 2017.

[12] Martin J Wainwright and Michael I Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, Jan. 2008.

[13] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge Univ. Pr, 2004.

[14] Ruslan Salakhutdinov, Sam Roweis, and Zoubin Ghahramani, "On the convergence of bound optimization algorithms," in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, 2003, UAI'03, pp. 509–516, Morgan Kaufmann Publishers Inc.

[15] David P Helmbold, Robert E Schapire, Yoram Singer, and Manfred K Warmuth, "A comparison of new and old algorithms for a mixture estimation problem," *Machine learning*, vol. 27, no. 1, pp. 97–119, 1997.

[16] Amir Globerson, Terry Y Koo, Xavier Carreras, and Michael Collins, "Exponentiated gradient algorithms for log-linear structured prediction," in *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA, 2007, ICML '07, pp. 305–312, ACM.

[17] Christopher M Bishop, *Pattern Recognition and Machine Learning*, Springer, 17 Aug. 2006.

[18] Masa-Aki Sato, "Online model selection based on the variational bayes," *Neural computation*, vol. 13, no. 7, pp. 1649–1681, 2001.