DAPPER: Scaling Dynamic Author Persona Topic Model to Billion Word Corpora

Robert Giaquinto
Dept. of Computer Science and Engineering
University of Minnesota
Twin Cities, USA
giaquinto.ra@gmail.com

Arindam Banerjee
Dept. of Computer Science and Engineering
University of Minnesota
Twin Cities, USA
banerjee@cs.umn.edu

***Abstract*—Extracting common narratives from multi-author dynamic text corpora requires complex models, such as the Dynamic Author Persona (DAP) topic model. However, such models are complex and can struggle to scale to large corpora, often because of challenging non-conjugate terms. To overcome such challenges, in this paper we adapt new ideas in approximate inference to the DAP model, resulting in the DAP Performed Exceedingly Rapidly (DAPPER) topic model. Specifically, we develop Conjugate-Computation Variational Inference (CVI) based variational Expectation-Maximization (EM) for learning the model, yielding fast, closed form updates for each document, replacing iterative optimization in earlier work. Our results show significant improvements in model fit and training time without needing to compromise the model's temporal structure or the application of Regularized Variation Inference (RVI). We demonstrate the scalability and effectiveness of the DAPPER model by extracting health journeys from the CaringBridge corpus — a collection of 9 million journals written by 200,000 authors during health crises.**

***Index Terms*—topic modeling, graphical model, regularized variational inference, healthcare, text mining, approximate inference, non-conjugate models**

I. Introduction

Topic modeling is a popular technique for automatically discovering compact, interpretable, latent representations of corpora. Many corpora exhibit an important structure, such as authorship or a temporal dependency between documents. Classic topic models like Latent Dirichlet Allocation (LDA) scale to large datasets [1]–[3], but do not account for any special structure in the corpus. Subsequent topic models are designed around such corpora, or are reparameterized to capture other features in the texts. For instance, the Correlated Topic Model (CTM) captures correlations between topics [4]. The added complexity of these models comes at a cost, however. In the case of CTM, the model is parameterized with non-conjugate terms — resulting in an additional variational parameter and requiring conjugate gradient descent to be run repeatedly on each document of the corpus. Up until recently CTM defined the standard approach in dealing with non-conjugate terms in variational inference. Topic models uniquely designed for corpora with a temporal structure, such as Dynamic Topic Model (DTM) and Continuous Time Dynamic Topic Model (CDTM), face similar issues as the CTM [5], [6]. In each of these models the scalability is compromised by non-conjugate terms.

In recent work, the Dynamic Author-Persona (DAP) topic model was introduced for corpora with multiple authors writing over time [7]. DAP represents each author by a latent persona — where personas capture the propensity to discuss certain topics over time. However, inference in DAP inherited the challenges with non-conjugacy from CTM and DTM.

In this paper, we seek to improve the scalability of the DAP topic model. Our approach is to adapt new ideas in approximate inference to DAP's variational Expectation-Maximization (EM) algorithm. Specifically, we develop a Conjugate-Computation Variation Inference (CVI) based variational EM algorithm, a powerful approach for transforming inference in non-conjugate models to conjugate models, leading to fast, closed form updates to parameters [8]. The advantage of CVI over other related approaches is that it preserves the closed form updates to parameters in the conjugate terms. We show how a CVI based inference algorithm applies to a complex, temporal topic model like DAP, and how this new inference algorithm improves model performance and dramatically reduces the time required to train the model.

Our primary motivation for developing a faster inference algorithm for the DAP model is the desire to scale the model to the CaringBridge (CB) corpus, which is a collection of 9 million journals ($\approx$1 billion words) written by approximately 200,000 authors during a health crisis. CaringBridge journals are written by patients and caregivers and posted to the CaringBridge website, to be shared privately with friends and family. The CB corpus holds enormous potential for insights on the challenges and experiences faced by those with serious, and often life threatening illnesses. The size and complexity of the data, however, present a modeling challenge too great, until now.

Our results show that the DAPPER model achieves likelihoods better than competing models, including LDA, DTM, CDTM, and DAP. Moreover, we show that DAPPER's conjugate-computation updates result in significant improvements in speed over its predecessor. Finally, we demonstrate the scalability of the DAPPER model by training it to the CB and Signal Media One-Million News Article corpora, and share the compelling narratives found by DAPPER's latent

personas.

The rest of the paper is as follows: in Section II, a background on recent advances in approximate inference is given. Section III presents a brief overview of the DAP model. Section IV details the CVI approach for accelerating the DAP model and describes a connection between CVI and expectation propagation. Section V introduces the evaluation datasets and procedures. Section VI shares the results of the experiments. Finally, Section VII summarizes the contributions of this paper.

II. Background

Approximate inference plays an important role in fitting complex probabilistic graphical models (PGM) which often have intractable posteriors and cannot be computed exactly. Interest in approximate inference techniques like variational inference, is growing because it tends to scale better than classical techniques, such as Markov Chain Monte Carlo [9]. Variational inference, in particular, transforms the inference problem into an optimization problem with the goal of finding hidden variables $\mathbf{z}$ to the variational distribution q such that $q(\mathbf{z})$ closely approximates the posterior $p(\mathbf{z} \mid \mathbf{y})$, where $\mathbf{y}$ are the observed data [10]. This equates to minimizing KL divergence between the approximate and true posterior:

$$q^*(\mathbf{z}) = \arg\min_{q \in \mathcal{Q}} KL(q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{y})) \ .$$

In PGMs the dependency between nodes, and their corresponding probability distributions, can form either conjugate or non-conjugate pairs. Non-conjugate pairs occur in a number of famous models, such as CTM and DTM [4], [5]. The challenge with non-conjugate priors, however, is that the posterior does not belong to the same family as the prior, and often the posterior cannot be obtained in a closed form analytically [11]. As a result, models with non-conjugate terms often require the introduction of additional variational parameters and gradient based optimizations which significantly slows down training.

Advances in Variational Inference In recent years tremendous progress has been made in improving both the speed, quality, and ease of application of variational inference. In 2013, Hoffman et al. introduced stochastic variational inference (SVI): a method that reparameterizes the gradient of the Expected Lower BOund (ELBO) in terms of the natural parameters in order to derive a fast stochastic gradient descent (SGD) algorithm for variational inference [2], [3]. The SVI approach is an important contribution because of the tremendous speed-up that results. Further, reparameterizing the ELBO so as to derive natural gradients, which leads to stochastic optimization, is closely related to traditional coordinate ascent variational inference. Natural gradients provide more stable learning, as opposed to gradient methods that are better suited for optimization in Euclidean geometry [12]. SVI is limited in that it requires the model's parameters to have an exponential family form, and hence is not directly applicable to non-conjugate models like CTM or DTM.

Recent advances, like Black Box Variational Inference (BBVI), demonstrate a promising new way to speed-up updates to non-conjugate terms [13]. The approach introduced in BBVI uses stochastic gradient updates, where the noisy stochastic observations are computed using Monte Carlo techniques. BBVI results in a variational inference algorithm that is faster than conventional approaches and eliminates the need to derive inference algorithms for new models.

Conjugate-Computation Variational Inference The computational downside of BBVI is that does not take advantage of conjugate terms with closed form updates. Khan and Lin introduce Conjugate-computation Variational Inference (CVI) which cleverly allows inference on models with non-conjugate terms to be computed as a *conjugate computation* [8]. A conjugate computation is simply the adding of the natural parameters of a prior to the sufficient statistics of the likelihood. In short, the CVI approach allows for fast updates to complex PGMs. Moreover, unlike SVI which takes gradients of the ELBO in the natural-parameter space, CVI uses stochastic mirror descent in the mean-parameter space that eschews Euclidean geometry (i.e. squared loss) in favor of a Bregman divergence defined by the convex-conjugate of the log-partition function. Khan and Lin [8] demonstrate that this approach leads to inference in a conjugate model, where non-conjugate terms have been replaced by exponential family approximations. Further, CVI lets models be trained with stochastic mini-batches — in the style of SVI. As a result, even complex models with difficult non-conjugate terms can be trained quickly and efficiently.

The goal of the CVI algorithm is to maximize a lower bound to the marginal likelihood:

$$\arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\lambda})]$$

where $\boldsymbol{\Lambda}$ is the set of valid variational parameters, $\boldsymbol{\lambda}$ the variational parameter, and $q(\mathbf{z} \mid \boldsymbol{\lambda})$ the variational approximation. Traditionally, the bound is optimized via gradient descent, i.e. $\boldsymbol{\lambda}_{i+1} \leftarrow \boldsymbol{\lambda}_{+} \rho_i \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_i)$, where ρ_i is the learning rate. An equivalent formulation of this gradient, which highlights the divergence function, is:

$$\boldsymbol{\lambda}_{i+1} \leftarrow \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \langle \boldsymbol{\lambda}, \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_i) \rangle - \frac{1}{2\rho_i} ||\boldsymbol{\lambda} - \boldsymbol{\lambda}_i||_2^2$$

CVI assumes distributions are minimal exponential families, meaning there is a one-to-one mapping between $\boldsymbol{\lambda}$ and the mean parameters $\boldsymbol{\mu} \in \mathcal{M}$. The lower bound is reparameterized in terms of $\boldsymbol{\mu}$ such that $\widetilde{\mathcal{L}}(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\lambda})$, and mirror descent gradient updates for this bound are derived. Additionally, in making the mean-field assumption, which assumes that the parameters are *posteriori* independent, the gradient update can be expressed as a summation over all nodes $k \in 1, \ldots, M$:

$$\max_{\boldsymbol{\mu}} \sum_{k=1}^{M} \left[\left\langle \boldsymbol{\mu}_k, \widehat{\nabla}_{\boldsymbol{\mu}} \widetilde{\mathcal{L}}(\boldsymbol{\mu}_i) \right\rangle - \frac{1}{\rho_i} \mathbf{B}_{A^*}(\boldsymbol{\mu}_k || \boldsymbol{\mu}_{k,i}) \right] , \quad (1)$$

where i refers to the iteration number, and $\mathbf{B}_{A^*}$ is a Bregman divergence — such as KL divergence, defined over the convex-conjugate of the log-partition A^*. The choice of divergence function is to account for the geometry of the parameter space. Khan et al. prove convergence for the general case of Bregman divergences, even in the stochastic gradient setting [14]. The maximization in (1) only requires optimizing for a single node k and hence can be done either in parallel, or as a doubly stochastic scheme by randomly picking a term in the summation.

One of the primary results proved by Khan and Lin [8] is that (1) can be implemented as Bayesian inference in a conjugate model. Their method hinges splitting the joint distribution into non-conjugate and conjugate terms (denoted $\tilde{p}_{nc}(\mathbf{y}, \mathbf{z})$ and $\tilde{p}_c(\mathbf{y}, \mathbf{z})$, respectively) and replacing the difficult non-conjugate term with an exponential family approximation whose natural parameter is $\widetilde{\boldsymbol{\lambda}}_i$. Hence, the posterior is approximated with a variational distribution defined by:

$$q(\mathbf{z} \mid \boldsymbol{\lambda}_{i+1}) \propto \exp(\phi(\mathbf{z}), \widetilde{\boldsymbol{\lambda}}_i) \widetilde{p}_c(\mathbf{y}, \mathbf{z}) \ ,$$

where $\widetilde{\boldsymbol{\lambda}}_i$ is the natural parameter of the exponential-family approximation to $\widetilde{p}_{nc}$, computed as a weighted sum of the gradients of the non-conjugate term. Khan and Lin [8] show that the exponential-family approximation's parameter $\widetilde{\boldsymbol{\lambda}}_i$ and the variational posterior's parameter $\boldsymbol{\lambda}$ are updated by:

$$\widetilde{\boldsymbol{\lambda}}_{k,t} = \sum_{a \in \mathbb{N}_k} \mathbb{E}_{q/k,i}[\boldsymbol{\eta}_{a,k}(\mathbf{z}_{a/k}, \mathbf{y}_{a/k})] + \nabla_{\mu_k} \mathbb{E}_{q_i}[\log \widetilde{p}_{nc}^{a,k}] \tag{2a}$$

$$\boldsymbol{\lambda}_{i+1} = (1 - \rho_i)\boldsymbol{\lambda}_i + \rho_i \widetilde{\boldsymbol{\lambda}}_i \tag{2b}$$

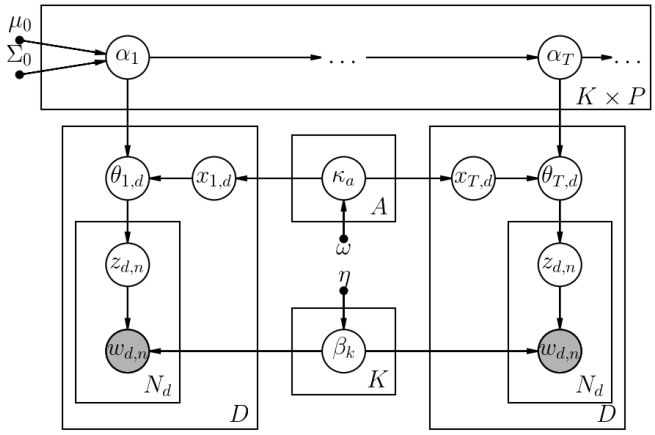
where $\boldsymbol{\eta}_{a,i}(\mathbf{z}_{a/i}, \mathbf{y}_{a/i})$ are simply the natural parameters for the conjugate parts of the model, and $\mathbb{N}_k$ the local neighborhood containing $\mathbf{z}_i$ and its children. By replacing non-conjugate terms with exponential family approximations, CVI allows even complex models to be trained quickly and efficiently.

III. Dynamic Author-Persona Topic Model

The Dynamic Author-Persona (DAP) topic model is designed for corpora with multiple authors writing over time [7]. Giaquinto et al. introduce the DAP model and demonstrate its ability to identify common narratives shared by patients and caregivers journaling during a serious health crisis on the website CaringBridge. While the model can produce valuable qualitative results from smaller datasets, it struggles to scale to industrial sized problems.

To model temporal dependencies between parameters DAP uses a Variational Kalman Filter, similar to [5], [6]. The structure of the DAP model, shown in Figure 1, is particularly unique due to the parameter $\boldsymbol{\alpha}_{t,p}$, which captures the distribution over topics for each persona p at time point t. The structure of the DAP model results in the joint distribution that factorizes as:

Fig. 1. Graphical representation of the Dynamic Author-Persona topic model (DAP). On top, topic distributions for each persona evolve over time by $\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1} \sim \mathcal{N}(\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Sigma})$. The distribution over words for each topic is $\boldsymbol{\beta} \sim Dir(\eta)$. Each author $a \in \{1, \ldots, A\}$ is represented by a distribution over personas defined by $\boldsymbol{\kappa}_a \sim Dir(\omega)$. The distribution over topics for each document $\boldsymbol{\theta}_d \sim \mathcal{N}(\boldsymbol{\alpha}_t \mathbf{x}_{t,d}, \boldsymbol{\Sigma}_t)$ is dependent on the persona assignment $\mathbf{x}_{t,d} \sim Mult(\boldsymbol{\kappa}_a)$ for that document's author, and the evolving topic distribution $\boldsymbol{\alpha}_t$. Words, denoted $\mathbf{w}$, are assigned to topics according to the multinomial $\mathbf{z}_{d,n} \sim Mult(\sigma(\boldsymbol{\theta}_{t,d}))$.



$$\prod_{k=1}^{K} p(\boldsymbol{\beta}_k \mid \eta) \prod_{a=1}^{A} p(\boldsymbol{\kappa}_a \mid \omega) \prod_{t=1}^{T} \prod_{p=1}^{P} p(\boldsymbol{\alpha}_{t,p} \mid \boldsymbol{\alpha}_{t-1,p}, \boldsymbol{\Sigma}_{t-1}) \times \prod_{d=1}^{D_t} p(\mathbf{x}_{t,d} \mid \boldsymbol{\kappa}_{a_d}) p(\boldsymbol{\theta}_{t,d} \mid \boldsymbol{\alpha}_{t,1:P} \mathbf{x}_{t,d}, \boldsymbol{\Sigma}_t) \prod_{n=1}^{N_{d_t}} p(\mathbf{z}_{d,n} \mid \sigma(\boldsymbol{\theta}_{t,d})) \ , \tag{3}$$

where $\sigma(\cdot)$ is a softmax function introduced to obey the constraint that $\mathbf{z}_{d,n}$ lies on the simplex.

The structure and parameterization of the model, however, introduces a number of non-conjugate terms, namely $p(\mathbf{z}_{d,n} \mid \sigma(\boldsymbol{\theta}_{t,d}))$ and $p(\boldsymbol{\theta}_{t,d} \mid \boldsymbol{\alpha}_t \mathbf{x}_{t,d}, \boldsymbol{\Sigma}_t)$. Consequently, estimating the topic assignment $\mathbf{z}$, topic proportions $\boldsymbol{\theta}$, and persona assignment $\mathbf{x}$ is challenging. The remaining model terms, the variational parameter used in the mean-field variational inference algorithm, and a brief description is given in Table I.

The DAP model's scalability issues stem from its non-conjugate terms — for which there are no fast, closed form updates. To derive parameter updates the DAP model's intractable posterior is approximated with a variational posterior under the mean-field assumption. In standard fashion, the Evidence Lower BOund (ELBO) is maximized, which is equivalent to minimizing the KL divergence between the variational in true posteriors. Once the ELBO is specified, updates are derived for each parameter by selecting terms containing that parameter and optimizing. However, due to the non-conjugate terms, fast, closed-form updates are not always possible. In particular the DAP model's E-Step — which runs multiple times for each document in the corpus — must use exponentiated gradient descent to learn the persona assignment $\boldsymbol{\tau}$ of an author, and conjugate-gradient descent to

TABLE I
NOTATION AND PARAMETERS USED IN THE DAP MODEL. VARIATIONAL REFERS TO THE CORRESPONDING PARAMETER IN THE MEAN-FIELD VARIATIONAL INFERENCE ALGORITHM.

Parameter	Variational	Description
$\mathbf{w}_{t,d}$		Words in document d_t
$\mathbf{z}_n$	$\boldsymbol{\phi}_n$	Assigns word n to a topic
$\boldsymbol{\theta}_{t,d}$	$\boldsymbol{\gamma}_{t,d}$	Topic distribution for document d_t
$\mathbf{v}_{t,d}$	$\hat{\mathbf{v}}_{t,d}$	Covariance between topics for d_t
μ_0		Prior for mean of $\boldsymbol{\alpha}_0$
Σ_0		Prior for covariance of $\boldsymbol{\alpha}_0$
$\boldsymbol{\alpha}_{t,p}$	$\hat{\boldsymbol{\alpha}}_{t,p}$	Persona p's topic distribution
$\boldsymbol{\Sigma}_t$	$\hat{\boldsymbol{\Sigma}}_t$	Covariance in topic distributions
ω		Prior for $\boldsymbol{\kappa}_a$
$\boldsymbol{\kappa}_a$	$\boldsymbol{\delta}_a$	Author a's personas distribution
$\mathbf{x}_{d,t}$	$\boldsymbol{\tau}_{t,d}$	Assigns author of d_t to a persona
η		Prior parameter for β_k
$\boldsymbol{\beta}_k$	$\boldsymbol{\lambda}_k$	$\forall k$ distribution over words

learn the mean and variance parameters of the document's topic distribution.

IV. DAP PERFORMED EXCEEDINGLY RAPIDLY

The non-conjugate terms in the DAP model compromise the scalability of the model. CVI presents an opportunity to directly address DAP's bottlenecks, while keeping the existing closed-form parameter updates. We refer to DAP trained with the new CVI based inference algorithm as Dynamic Author-Persona Performed Exceedingly Rapidly (DAPPER). Details of the derivation of DAPPER's inference algorithm are shared below. In particular, the algorithm is structured like variational EM, where local variational parameters are updated in the Expectation step on a mini-batch or the entire corpus, and then global model parameters are updated in the Maximization step.

A. E-Step

In the E-step we update each document d's topic proportions $\boldsymbol{\theta}_d$, the assignment of each word to a topic $\mathbf{z}_n$, and the assignment of each author to a persona $\mathbf{x}_d$.

1) Document Topic Proportions: Each document is given a hidden parameter $\boldsymbol{\theta}_d$ representing the proportions of each topic. We begin by identifying the conjugate and non-conjugate terms involving $\boldsymbol{\theta}_d$. For $\boldsymbol{\theta}_d$ we have a conjugate term: $\tilde{p}_c^{\theta} = \mathcal{N}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_t \mathbf{x}_d, \Sigma_t)$. Here $\mathcal{N}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_t \mathbf{x}_d, \Sigma_t)$ is a Gaussian conditioned on a Gaussian because the mean parameter is the dot product between two fixed terms $\boldsymbol{\alpha}_t \mathbf{x}_d$, giving a natural parameter:

$$\begin{bmatrix} \Sigma_t^{-1}(\boldsymbol{\alpha}_t \mathbf{x}) & -\frac{1}{2}\Sigma_t^{-1} \end{bmatrix}^\top \tag{4}$$

The second term involving $\boldsymbol{\theta}$ required is the non-conjugate term, which is $\tilde{p}_{nc}^{\theta} = Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}_d))$

The variational distribution is defined $q(\theta_{d,k}) = \mathcal{N}(\theta_{d,k} \mid \gamma_{d,k})$ where $\gamma_{d,k} = \{m_k, v_k\}$ for topic k has sufficient statistics:

$$ss(\boldsymbol{\theta}_{d,k}) = \begin{bmatrix} \theta_{d,k} & \theta_{d,k}^2 \end{bmatrix}^\top$$

Writing the approximate posterior $q(\boldsymbol{\theta})$ as a product of the conjugate and non-conjugate parts gives:

$$q_{i+1}(\boldsymbol{\theta}) \propto \Big[\prod_{k=1}^{K} \exp(ss(\theta_{d,k}) \widetilde{\boldsymbol{\Theta}}_{k,i}) \Big] \mathcal{N}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_t \mathbf{x}_d, \Sigma_t)$$

where $\widetilde{\boldsymbol{\Theta}}_{k,i}$ are the natural parameters to the approximated exponential family at iteration i.

The difficult non-conjugate term $Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))$ has already been approximated in [4], specifically:

$$\begin{aligned} f &= \mathbb{E}_q \Big[\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta})) \Big] \\ &\geq \sum_{k=1}^{K} m_k \phi_n^{(k)} - \zeta^{-1} \sum_{k=1}^{K} \exp(m_k + \frac{v_k}{2}) - \log(\zeta) + 1 \end{aligned} \tag{5}$$

where, again, the parameter ζ is introduced to preserve a lower bound. Thus, we can now take the gradient of $f = \mathbb{E}_q[\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))]$ with respect to the mean parameters.

$$\nabla_\mu f = \begin{bmatrix} \frac{\partial f}{\partial \mu^{(1)}} & \frac{\partial f}{\partial \mu^{(2)}} \end{bmatrix}^\top$$

Since the mean variational distribution's mean parameters are $\boldsymbol{\mu} = \begin{bmatrix} m_k & m_k^2 + v_k \end{bmatrix}^\top$, we can write $m_k = \mu_k^{(1)}$ and $v_k = \mu_k^{(2)} - (\mu_k^{(1)})^2$. By the chain rule the gradient with respect to the mean parameters are $\frac{\partial f}{\partial \mu^{(1)}} = \frac{\partial f}{\partial m} - 2\frac{\partial f}{\partial v}$ and $\frac{\partial f}{\partial \mu^{(2)}} = \frac{\partial f}{\partial v}$. Applying these gradients to the non-conjugate term in (5) we can then compute the natural parameter of the variational posterior:

$$\begin{aligned} \frac{\partial f}{\partial m} - 2\frac{\partial f}{\partial v} m = \phi_n^{(k)} \quad \text{and} \quad \frac{\partial f}{\partial v_k} = \frac{1}{2\zeta} \exp(m_k + \frac{v_k}{2}) \\ \implies \nabla_{\boldsymbol{\mu}}(f) = \begin{bmatrix} \phi_n^{(k)} & -\frac{1}{2\zeta} \exp(m_k + \frac{v_k}{2}) \end{bmatrix}^\top \end{aligned} \tag{6}$$

where ζ has the same update as before: $\zeta \leftarrow \sum_{k=1}^{K} \exp(m_k + \frac{v_k}{2})$. Thus by the CVI update rules given in (2), the natural parameter of the topic proportions is given by a *conjugate computation* adding (6) (the sufficient statistics) to (4) (the natural parameter of prior):

$$\boldsymbol{\Theta}_{k,i+1} = \rho_i \begin{bmatrix} \sum_{n=1}^{N_d} \phi_n^{(k)} + \Sigma_t^{-1}(\boldsymbol{\alpha}_t \mathbf{x}_d)_k \\ \frac{-N}{2\zeta} \exp(m_k + \frac{v_k}{2}) + (-\frac{1}{2}\Sigma_{t,k,k}^{-1}) \end{bmatrix} + (1 - \rho_i)\boldsymbol{\Theta}_{k,i} \tag{7}$$

Ideally we want to compute the source parameters to the variational posterior, i.e. m_k and v_k — which is straightforward[1] given the natural parameter $\boldsymbol{\Theta}_{k,i+1}$ computed in (7). The final updates to the variational posterior's source mean and variances are computed:

$$m_{k,i+1} = \frac{-\Theta_{k,i+1}^{(1)}}{2\Theta_{k,i+1}^{(2)}} \quad \text{and} \quad v_{k,i+1} = \frac{-1}{2\Theta_{k,i+1}^{(2)}}$$

[1]These follow from the definitions for converting a Multivariate Gaussian between its source and natural parameters, which can easily be looked up, see for example [15].

2) Topic Assignment: Each word is assigned to a topic through a hidden parameter $\mathbf{z}_n$. For $\mathbf{z}_n$ the corresponding conjugate term is $\tilde{p}_c^{\mathbf{z}} = Mult(\mathbf{w}_n \mid \boldsymbol{\beta}_{z_n})$, and non-conjugate term is $\tilde{p}_{nc}^{\mathbf{z}} = Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))$.

Define the variational distribution $q(\mathbf{z}_n) = Mult(\mathbf{z}_n \mid \phi_n)$ for word n. Writing the approximate posterior $q(\mathbf{z}_n)$ as a product of the conjugate and non-conjugate parts:

$$q_{i+1}(\mathbf{z}_n) \propto \Big[\prod_{n=1}^{N_d} \exp(ss(\mathbf{z}_n)\widetilde{\boldsymbol{\Phi}}_i)\Big] Mult(\mathbf{w}_n \mid \boldsymbol{\beta}_{z_n}) \tag{8}$$

where $\widetilde{\boldsymbol{\Phi}}_i$ are the natural parameters to the approximated exponential family at iteration i, computed by:

$$\begin{aligned}\widetilde{\boldsymbol{\Phi}}_i &= \Big[\widetilde{\Phi}_{1,i} \quad \dots \quad \widetilde{\Phi}_{K,i}\Big]^\top \\ &= \rho_i \nabla_\phi \mathbb{E}_{q_i}\Big[\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))\Big]|_{\phi=\phi_i} + (1-\rho_i)\widetilde{\boldsymbol{\Phi}}_{i-1}\end{aligned}$$

Using the previously computed approximation to the challenging term $\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))$ in (6), we now differentiate it with respect to the mean parameter of $q(\mathbf{z}_n)$, i.e. ϕ. This gives:

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q_i}\Big[\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))\Big] = \begin{bmatrix} m_1 & \dots & \dots m_K \end{bmatrix}^\top = \mathbf{m} \tag{9}$$

where m_k is the mean of $q(\theta_{d,k}) = \mathcal{N}(\theta_{d,k} \mid m_k, v_k)$, for $k \in 1, \dots, K$. To update the natural parameter to the approximated exponential family in (8), we use the equations:

$$\begin{aligned}\widetilde{\boldsymbol{\Phi}}_i &= \rho_i\Big(\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q_i}[\log Mult(\mathbf{z}_n \mid \sigma(\boldsymbol{\theta}))]\Big) + (1-\rho_i)\widetilde{\boldsymbol{\Phi}}_{i-1} \\ &= \rho_i \mathbf{m} + (1-\rho_i)\widetilde{\boldsymbol{\Phi}}_{i-1}\end{aligned} \tag{10}$$

Since CVI transforms non-conjugate computations into conjugate computations the update to the variational parameter $\boldsymbol{\phi}_n$ is similar to the LDA case. Specifically, we compute the source parameter of our variational posterior $\boldsymbol{\phi}_n$ by:

$$\phi_{n,k,i+1} \propto \exp(\widetilde{\Phi}_{k,i} + \mathbb{E}_q[\log \beta_{k,w_n}]) \tag{11}$$

where, as usual, the Dirichlet expectation $\mathbb{E}_q[\log \beta_{k,v}]$, is computed by $\Psi(\lambda_{k,v}) - \Psi(\sum_{j=1}^{V} \lambda_{k,j})$. The first term $\widetilde{\Phi}_{k,i}$ is the document's topic distribution computed in the previous section, hence (as expected) the CVI update for $\boldsymbol{\phi}$ results in a simple closed form update with the same form as in the original DAP model.

3) Persona Assignment: For each document the author is assigned to a persona through a hidden parameter $\mathbf{x}_d$. For $\mathbf{x}_d$ the conjugate term is $\tilde{p}_c^{\mathbf{x}} = Mult(\mathbf{x}_a \mid \kappa_{d_a})$, and its non-conjugate term is $\tilde{p}_{nc}^{\mathbf{x}} = \mathcal{N}(\boldsymbol{\theta} \mid \alpha_t \mathbf{x}_d, \Sigma_t)$, where the coupling in the mean $\alpha_t \mathbf{x}_d$ made closed form updates in the DAP model impossible.

Define the variational distribution $q(\mathbf{x}_d) = Mult(\mathbf{x}_d \mid \tau_d)$. Writing the approximate posterior $q(\mathbf{x}_d)$ as a product of the conjugate and non-conjugate parts gives:

$$q_{i+1}(\mathbf{x}_d) \propto \Big[\prod_{p=1}^{P} \exp(ss(\mathbf{x}_d)\widetilde{\tau}_{d,p,i})\Big] Mult(\mathbf{x}_d \mid \boldsymbol{\kappa}_{d_a}) \tag{12}$$

where $\widetilde{\boldsymbol{\tau}}_{d,i}$ are the natural parameters to the approximated exponential family at iteration i.

We compute the variational posterior by first taking the gradient of the non-conjugate terms, where the non-conjugate term $f = \mathbb{E}_{q_i}[\mathcal{N}(\boldsymbol{\theta} \mid \alpha_t \mathbf{x}_d, \Sigma_t)]$ evaluates to:

$$\begin{aligned}f = \frac{-1}{2}\Big((\gamma_{t,d} - \hat{\alpha}_t \tau_{t,d})^\top \Sigma_t^{-1}(\gamma_{t,d} - \hat{\alpha}_t \tau_{t,d}) + \\ \sum_{p=1}^{P} \mathrm{Tr}\left[\Sigma_t^{-1} \operatorname{diag}\left(\tau_{t,d,p}(\hat{\alpha}_{t,p}\hat{\alpha}_{t,p}^\top + \hat{\Sigma}_t)\right)\right]\Big) + const\end{aligned}$$

Taking the gradient with respect to the mean parameter $\boldsymbol{\tau}$ gives:

$$\nabla_\tau f = \hat{\alpha}_{t,p}\Sigma_t^{-1}(\gamma_{t,d} - \hat{\alpha}_{t,p}\tau_{t,d,p}) - \frac{1}{2}\mathrm{Tr}(\Sigma_t^{-1} \operatorname{diag}(\hat{\alpha}_{t,p}^2 + \hat{\Sigma}_t))$$

Next we use the CVI updates rule from (2) to compute the natural parameter of our variational posterior $q(\mathbf{x}_d)$ by:

$$\widetilde{\boldsymbol{\tau}}_{d,i} = \rho_i\Big(\mathbb{E}_q[\log \boldsymbol{\kappa}_{d_a}] + \nabla_\tau f\Big) + (1-\rho_i)\widetilde{\boldsymbol{\tau}}_{d,i-1}$$

where $\mathbb{E}_q[\log \boldsymbol{\kappa}_{d_a}]$ is the expected natural parameters from the conjugate term, and is equivalent to a Dirichlet expectation: $\Psi(\delta_{a,p}) - \Psi(\sum_{j=1}^{P} \delta_{a,j})$. In order to map the natural parameter $\widetilde{\boldsymbol{\tau}}_{d,i}$ back to the source parameter of the variational posterior $q(\mathbf{x}_d)$, we use

$$\boldsymbol{\tau}_{d,i} = \begin{bmatrix} \frac{\exp(\widetilde{\boldsymbol{\tau}}_{d,1,i})}{\sum_{p=1}^{P} \widetilde{\boldsymbol{\tau}}_{d,p,i}} & \dots & \frac{\exp(\widetilde{\boldsymbol{\tau}}_{d,P,i})}{\sum_{p=1}^{P} \widetilde{\boldsymbol{\tau}}_{d,p,i}} \end{bmatrix}^\top$$

B. M-Step

In the M-step we use sufficient statistics collected from computing document-level variational parameters computed during the E-step to update the global parameters $\boldsymbol{\beta}, \boldsymbol{\kappa}$, and $\boldsymbol{\alpha}$. Because DAPPER makes use of stochastic mini-batches, we use the learning rate defined for SVI and recommended in CVI: $\rho_i = (i+\tau)^{-\kappa}$ where $\tau \geq 0$ is the delay and $\kappa \in (0.5, 1.0]$ is the forgetting rate [2], [3], [8].

1) Topic's Distribution over Words: The $\boldsymbol{\beta}$ term is already conjugate, and hence the variational distribution for the topics, $q_{i+1}(\boldsymbol{\beta}) = \prod_{k=1}^{K} Dir(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_{k,i+1})$, already has a closed form update: $\boldsymbol{\lambda}_{k,i+1} = (1-\rho_i)\boldsymbol{\lambda}_{k,i} + \rho_i(\eta + \sum_{d=1}^{D}\sum_{n=1}^{N_d} \phi_{d,n,k} w_{d,n})$.

2) Author's Distribution over Personas: Since the $\boldsymbol{\kappa}$ terms are already conjugate and have closed form solutions, it follows that the update to variational posterior, $q_{i+1}(\boldsymbol{\kappa}) = \prod_{d=1}^{D} Dir(\boldsymbol{\kappa}_{d_a} \mid \boldsymbol{\delta}_{d_a,i+1})$, has a simple closed form solution using the convex combination: $\boldsymbol{\delta}_{d_a,i+1} = (1-\rho_i)\boldsymbol{\delta}_{d_a,i} + \rho_i(\omega + \sum_{d=1}^{D} \boldsymbol{\tau}_d)$.

3) Persona's Distribution over Topics: The $\boldsymbol{\alpha}_t$ term is conjugate to all other factors, and is global. The variational distribution for the distribution over topics for each persona $\boldsymbol{\alpha}_{t,1:P}$ is $q_{i+1}(\boldsymbol{\alpha}_{t,p}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{\alpha}_{t,p} \mid \hat{\boldsymbol{\alpha}}_{t-1,p,i+1}, \Sigma_t) \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{\theta}_d \mid m_d, v_d)$. As shown in the original derivation of the DAP model, a closed form update can be found for $\hat{\boldsymbol{\alpha}}_{t,p}$:

$$\hat{\boldsymbol{\alpha}}_{t,p}^{new} = \frac{\hat{\boldsymbol{\alpha}}_{t-1,p} + \sum_{d=1}^{D_t} \boldsymbol{\gamma}_{t,d} \tau_{t,d,p} - \sum_{d=1}^{D_t} \tau_{t,d,p}}{1 + \sum_{d=1}^{D_t} \tau_{t,d,p}^2} \tag{13}$$

Thus, for mini-batch training we update $\hat{\boldsymbol{\alpha}}_{t,p,i+1}$, by first computing $\hat{\boldsymbol{\alpha}}_{t,p}^{*}$ from a mini-batch of documents. Then $\hat{\boldsymbol{\alpha}}_{t,p,i+1}$ is updated via a convex combination: $\hat{\boldsymbol{\alpha}}_{t,p,i+1} = (1 - \rho_i)\hat{\boldsymbol{\alpha}}_{t,p,i} + \rho_i \hat{\boldsymbol{\alpha}}_{t,p}^{*}$. Alternatively, to encourage personas to be distinct the update (13) is replaceable by the Regularized Variational Inference (RVI) update for $\boldsymbol{\alpha}_{t,p}$ given in the DAP model. CVI compliments RVI because closed form updates, such as the $\hat{\boldsymbol{\alpha}}_{t,p}$ update found by the regularized DAP model, are preserved. After computing $\hat{\boldsymbol{\alpha}}_{t,p,i+1}$, we proceed as usual and apply the forward and backward equations of the variational Kalman Filter to smooth over time time steps.

C. Connect Between CVI and Expectation Propagation

While DAPPER's inference algorithm is based on CVI — a recent advance in approximate inference, CVI itself has theoretical connections to the well known expectation propagation (EP) algorithm [16], [17]. The EP algorithm, which is an extension of Assumed Density Filtering, infers the approximate posterior q using localized inferences. With posterior p, hidden parameters $\mathbf{z}$, and observations $\mathbf{y}$, we assume p can be written as a product of terms: $p(\mathbf{z} \mid \mathbf{y}) \propto \prod_{i=0}^{N} f_i(\mathbf{z})$, where $f_0(\mathbf{z}) = p(\mathbf{z})$ expresses the prior density and $f_i(\mathbf{z}) = p(\mathbf{y} \mid \mathbf{z})$ the likelihood. The EP algorithm then approximates the posterior, choosing an approximating family with density $q(\mathbf{z}) \propto \prod_{i=1}^{N} q_i(\mathbf{z})$ and iteratively incorporating $q_i(\mathbf{z})$ into $q(\mathbf{z})$. First, EP computes the cavity distribution — that is, deleting $q_i(\mathbf{z})$ from $q(\mathbf{z})$, by $q_{-i}(\mathbf{z}) \propto q(\mathbf{z})/q_i(\mathbf{z})$. Second, a true Bayesian update incorporates $f_i(\mathbf{z})$:

$$\hat{p}(\mathbf{z}) = Z_i^{-1} q_{-i}(\mathbf{z}) f_i(\mathbf{z}) \ , \quad Z_i = \mathbb{E}_{\mathbf{z} \sim q_{-i}}[f_i(\mathbf{z})]$$

where $\hat{p}(\mathbf{z})$ is a tilted exponential family. The exact posterior is approximated, for exponential families minimizing KL divergence between the posteriors,

$$q^{new}(\mathbf{z}) = \arg\min KL(\hat{p}(\mathbf{z}) || q(\mathbf{z})) \ , \tag{14}$$

corresponds to matching the moments of p and q. Finally, update $q_i^{new}(\mathbf{z})$ by $q_i^{new}(\mathbf{z}) \propto q(\mathbf{z}) q_{-i}(\mathbf{z})$. Note, that the update to $q_i(\mathbf{z})$ is the local minimization and can be formulated as:

$$q_i^{new}(\mathbf{z}) = \arg\min KL(f_i(\mathbf{z}) q_{-i}(\mathbf{z}) || q_i(\mathbf{z}) q_{-i}(\mathbf{z})).$$

From here two connections to CVI appear. First, EP also computes an exponential family approximation in its approximation of $\hat{p}(\mathbf{z})$, the tilted distribution induced by $f_i(\mathbf{z})$ [18]. While EP does this computation using moment matching, moment matching corresponds to minimizing the Kullback-Leibler divergence from the tilted distribution to the new approximated marginal distribution [19], and moments can be computed as derivatives of the log normalizer, hence:

$$\begin{aligned} \mu^{new} = \mathbb{E}_{\hat{P}}[\phi(\mathbf{z})] &= \nabla_{q_{-i}} \log Z_i + \mu_{-i} \\ &= \nabla_{q_{-i}} \log \mathbb{E}_{\mathbf{z} \sim q_{-i}}[f_i(\mathbf{z})] + \eta_{-i} \end{aligned}$$

For exponential families $\nabla \log Z = E[\phi(\mathbf{y})]$, and therefore this moment matching can be viewed as similar to conjugate computations, here we add expectations of sufficient statistics of q to the corresponding expectations of $\mathbf{z}$ in $q_{-i}(\mathbf{z}) f_i(\mathbf{z})$. This shows that EP's creation of the tilted distribution moment parameter is analogous to CVI's computation of the natural parameter to the exponential family approximation, i.e. (2a).

The second connection to CVI lies in the "damping" technique used to improve the convergence of EP. Damping replaces the generic update $\lambda_i \rightarrow \lambda_{i+1}$ by a convex combination, which reduces the step size so that only a partial update is applied. CVI also updates the natural parameters with a convex combination: $\lambda_{t+1} = (1 - \beta)\lambda_t + \beta \widetilde{\lambda}_t$ in CVI is essentially just "damped" updates in EP [19]. This form of updating is analogous to minimizing an alpha divergence (which includes directed KL as a special case).

Despite a number of similarities, EP and CVI differ critically in convergence guarantees. Khan et al. show that CVI converges under fairly mild assumptions, namely that q is a minimal exponential family and the model's conditional distribution can be split into conjugate and non-conjugate terms. EP, on the other hand, is not guaranteed to converge. EP minimizes KL divergence for each local observation, but does not directly minimize $KL(p \,||\, q)$.

V. Experiments

To evaluate the performance of DAPPER we perform a quantitative comparison with similar topic models (LDA, DTM, CDTM, and DAP), and a qualitative demonstration of DAPPER's scalability and output on the CB[2] and Signal Media One-Million News Article[3] (SM) corpora [21]. For the quantitative comparison per-word log-likelihoods ($PWLL$) are computed on test data, where $PWLL = \frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}$. While $PWLL$s do not correlate with a model's ability to discover coherent topics [22], they do offer a fair comparison of how well each model optimizes its objective function. Additionally, the speed and efficiency of DAPPER relative to its predecessor are measured by showing model performance as a function of training time. The qualitative comparison demonstrates the

[2]CB data were acquired with the permission and collaboration of CB leadership in accordance with CB's Privacy Policy & Terms of Use Agreement. Because of their highly sensitive content the CB dataset has been anonymized, but deidentification techniques are imperfect [20] and hence we cannot publicly release the CB dataset. Those interested in the dataset are encouraged to contact the investigators. All code for training the DAPPER model and running our experiments on the SM dataset, however, are available at https://github.com/robert-giaquinto/dapper.

[3]http://research.signalmedia.co/newsir16/signal-dataset.html

rich and compelling "health journeys" discovered by DAPPER on the CB corpus.

A. Datasets

Both the CB and SM corpora are pre-processed by removing common stopwords and reducing words to their lemma forms. Document timestamps are converted into a continuous, relative measure; for CB we use the number of weeks since author's first post (only looking at the first year of each authors journals), and for SM we use the day within span of the corpus (1-30 September, 2015).

CaringBridge. Established in 1997, CaringBridge is a 501(c)(3) non-profit organization that connects people and reduces the feelings of isolation that are often associated with a patient's health journey. DAPPER and its predecessor are designed with the CaringBridge corpus in mind. We demonstrate scalability and quality of DAPPER's output on the full CB corpus. The CaringBridge corpus consists of 9,010,623 journals written by 200,388 authors (with a total of 937,503,945 words) between 2006 and 2016 on the CaringBridge website. On average, authors write 100 words per journal and 45 journal posts in the first year.

For a qualitative evaluation, 22,552 randomly selected CB journals are set aside as a test set to evaluate the model and track convergence, leaving 8,988,071 journals in the training set. Our goal in training the DAPPER model on this dataset is to show that the model can find compelling qualitative results even on massive, complex datasets.

A quantitative evaluation on a subset of CB journals is drawn from 2,000 randomly selected authors, leaving a total of 114,532 journals. We refer to this corpus as CB-subset. From here 90% of the journals ($N = 103,018$) are divided into the training set, and the remaining 10% of each author's journals ($N = 11,728$) make up the test set. Training and test sets contain the same authors because personas distributions are learned for each author during training. These authors journal an average of 57 times during the first year, with a mean of 5 days between journal posts.

Signal Media Blogs. From the SM dataset we only consider articles written by bloggers who wrote fewer than one blog post per day during the corpus' one month span. Subsetting the data in this way is done to exclude major news organizations and instead focus on bloggers who typically write about a central theme. We refer to the subset as SM-blogs. After preprocessing, the SM-blogs corpus consists of 97,839 documents for training (15,848 blogs, 19,278,689 total words), and 10,887 documents for testing (same authors, 2,165,634 words).

B. Hyperparameters

To ensure a fair comparison we fix hyperparameters appearing in each of the models, such as number of topics and convergence criteria. Relative differences between model performances don't vary significantly depending on the number of topics chosen, and hence we only report results for models with 25 topics on the CB-subset and 50 topics on the SM-blogs. DAP and DAPPER seek 15 and 25 latent personas for the CB-subset and SM-blogs corpora, respectively, and fix their regularization of personas to $\rho = 0.2$. Because DAPPER can be trained on stochastic mini-batches we report results for various mini-batch sizes and full-batch training.

TABLE II
OVERALL COMPARISON OF MODELS AFTER A MAXIMUM OF 24 HOURS OF TRAINING ON CB-SUBSET AND SM-BLOGS CORPORA. PER-WORD LOG-LIKELIHOODS ARE REPORTED FOR THE TEST CORPUS.

Model	CB-subset	SM-blogs
DAPPER (full batch)	-6.73	-5.76
DAPPER (batch size = 512)	-8.19	-6.31
DAP	-8.84	-7.50
CDTM	-8.81	-8.24
DTM	-9.59	-7.93
LDA	-9.23	-7.79

TABLE III
HOURS OF TRAINING FOR DAPPER TO OUTPERFORM DAP'S BEST PERFORMANCE ($PWLL = -8.65$) ON THE CB-SUBSET TEST SET.

Batch Size	Hours to Exceed $PWLL = -8.65$	Speedup
DAP	40.43	Baseline
256	37.32	1.1x
512	2.21	18.3x
1024	1.97	20.5x
2048	2.13	19.0x
4096	4.11	9.8x
Full Batch	7.18	5.6x

VI. RESULTS

A. Model Performance Comparison

To compare the performance of DAPPER, we train and test DAPPER along with four similar topic models (LDA, DTM, CDTM, and DAP) on the quantitative corpora (see Section V). Each model is trained for a maximum of 24 hours on a single Haswell E5-2680v3 processor or until training performance converged — although the DAP model is the only model not to converge within 24 hours. Performance of each model is shown in Table II. The DAPPER model shows significant performance improvements over all competing models due to its faster method for handling non-conjugate terms. Performance for DAPPER is shown for a mini-batch size of 512, which consistently achieved the best training set performance after 24 hours, and DAPPER trained with full batch gradients updates, which achieved the best overall test performance after 24 hours.

Table II highlights three important results: first, the DAP topic model which is designed for the multi-author, temporal structure of the CB dataset achieves competitive performance but clearly suffers by not converging within 24 hours. Second, the DAPPER model benefits from faster training and achieves state-of-the-art performance. Third, smaller mini-batches like 512 result in good training performance that converges quickly but the model does not generalize as well as DAPPER trained with full batch gradients.

In Table IV we show the performance of the DAPPER model on the SM-blogs corpus with varying hyperparameter

settings. Specifically, we train models with [100, 75, 50, 25] topics, [50, 25, 15] latent personas, and mini-batch sizes of either [256, 512, 1024, 2048] or full gradient training. Full batch training results in the highest per-word log-likelihoods on the test set. Varying the number of personas and the number of topics has a noticable impact on performance (smaller models tend to do slightly better), however batch size is the most significant factor in achieving optimal performance. Smaller models (in terms of number of personas and topics) tend to do well on the 97,839 document SM-blogs corpus, however the best models used full batch training with 50 topics and either 15 or 25 personas.

B. Speed and Efficiency

Test set performance of the DAPPER model varies significantly depending on the batch size. Shown in Figure 2 is the performance of the DAP and DAPPER models trained on the CB-subset corpus and evaluated on the training and test sets after each epoch (one full pass over the training corpus). Each epoch of the DAP model takes an average of 6.7 hours to complete, whereas the DAPPER takes roughly 0.2 hours. The right plot (training set performance) shows that all DAPPER batch sizes begin to converge to a similar value. The training results (left plot in in Figure 2) show that all batch sizes converge to a similar value. On the test set (right plot in in Figure 2), however, larger batch sizes show better generalization.

Smaller batch sizes improve quickly at first but ultimately converge to lower PWLLs. The poor performance of small batch sizes may be due to the mini-batches being too noisy. Conversely, the larger batch sizes achieve the best performances, but improve slowly *at first*. We summarize this phenomenon in Table III, which reports how quickly DAPPER overtakes the optimal test set performance achieved by DAP. For example, a batch size of 256 converges almost immediately and takes many hours to eventually surpass DAP's best test set result. Whereas a batch size of 1024 improves steadily, and surpasses DAP's best PWLL in a fraction of the time. Despite the implication that the high variance of smaller batch sizes limits performance, we saw no benefit to gradient smoothing techniques, such as those proposed in [23].

C. Scalability and Qualitative Results

To demonstrate the scalability of the DAPPER model, we train DAPPER on the full CB corpus. Figure 3 presents selected personas discovered by a DAPPER model with 100 topics and 50 personas. The model is trained using a 24 processor machine for 94 hours, using a regularization of $\rho = 0.15$ and a batch size of 4096. DAPPER's efficient inference algorithm scales to massive datasets. Additionally, with stochastic updates only a constant amount of memory is required. The personas shown in Figure 3 highlight a variety of health journeys experienced by CB authors. In Table V we list the most likely words associated with each topic as well as the hand-defined labels assigned to each topic.

Scaling DAPPER to the full CB corpus makes it possible to build larger, richer models — which in turn can discover a broader range of narratives. Compared to results found by DAP in [7], DAPPER's scalability leads to the discovery of many new topics and personas. Many of the new topics discovered by DAPPER are unrelated health conditions. For example, "Friend, Memories," and "Life and Death" highlight how authors blend health and life updates in their journaling. This makes sense, a patient's condition is often known by readers or has been previously publicized on the author's homepage, and thus the focus of journals is instead on the patient's current health state. Moreover, health updates tend to focus on procedures (like medical tests and tools), or more general health descriptions like side-effects, infection, pain, or specific body-parts.

Finally, in Figure 4 we share qualitative results from DAPPER on the SM-blogs when trained with full batch gradients, 50 topics, and 15 personas — which was found to perform best during compared to other hyperparameter settings. Figure 4 shows the top three weighted topics for selected personas, highlighting how DAPPER discovers groups of authors whose writing blends unique topics over time. For instance, authors in Persona 12 tend to talk predominantly about the law and police in additional to reports and public records. Like a number of other personas, Persona 12 often references social media, which is associated with discussing something the author discovered through social media or the author encouraging readers to share and comment on their blog.

VII. Conclusion

While the structure of DAPPER mimics its predecessor, we derive a fundamentally new inference algorithm based on CVI. DAPPER surpasses its predecessor in terms of speed (35x faster), memory (constant requirements for mini-batch training), and significantly better likelihoods. DAPPER scales to massive datasets on commodity hardware, which in turn allows for deeper insights into topics, and common narratives hidden in the data. Additionally, we show that Regularized Variational Inference, which is applied to the DAPPER model to encourage distrinct personas, integrates with CVI cleanly because CVI preserves closed form updates. The success of DAPPER demonstrates that CVI can be applied to complex, temporal graphical models — eliminating the need to run multiple optimization procedures on each document, and instead replace all parameter updates with fast, closed form updates and stochastic mini-batch training.

While the work presented here demonstrates the DAPPER topic model's readiness for industrial-sized problems, there exist opportunities for further research. For one, our results show that too noisy of updates resulting from small mini-batches lead to poor performance. However, as briefly mentioned in the results, simple attempts to reduce variance through gradient averaging did not yield performance improvements. Further research is needed to find gradient updates that improve model performance in early iterations (as with small to medium batch sizes), but converge to better PWLLs (as with the larger batch

Fig. 2. Per-word Log-Likelihood performance on the CB training and test sets (larger is better). Each point represents the performance evaluated at the end of an epoch. Each model was trained for a maximum of 48 hours. DAPPER, which incorporates stochastic CVI updates, achieves better likelihoods and converges faster than the DAP model trained with variational EM. Performance of the DAPPER model varies by mini-batch size.

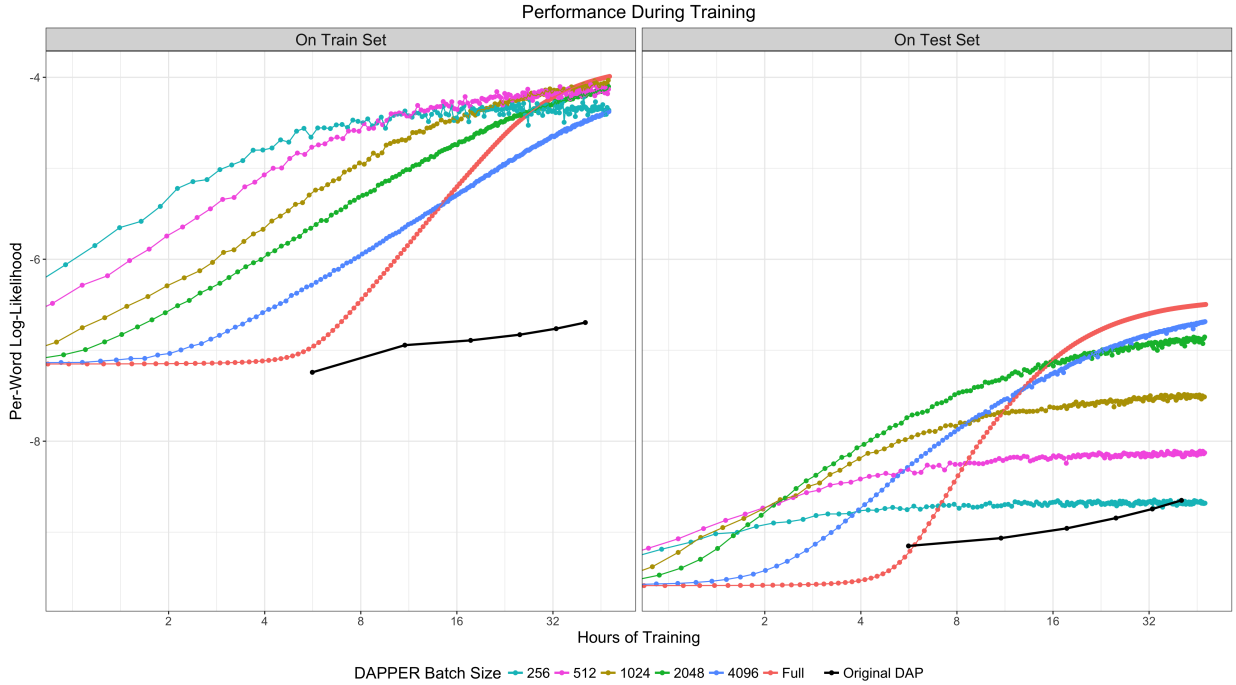


Fig. 3. Selected personas learned by the DAPPER model on the full CaringBridge collection of journals. Each plot shows a different persona, and the three topics most strongly associated with that persona. For clarity, topic labels are hand-defined based on the top words in the topic and journals most associated with that topic (see Table V). Personas show a variety of health journeys. An appeal to a higher power and prayer are common in many journals, and appear in personas 0, 29, and 42. Similarly, a deep reflection on life and death, possibly with respect to one's child appear in 0 and 29. Persona 22 captures a common experience of caring for an aging parent, beginning with intensive care and possibly ending with a hospice or nursing home. Persona 26 shows alternating periods of medical tests and intensive care with times of celebration. Personas 42 and 48 are both associated with cancer, but display very different narratives. Persona 48 includes the pair of topics "Insurance" and "URL Donation" which often appear together, indicating an author struggling with insurance and medical bills and seeking financial support from friends and family.

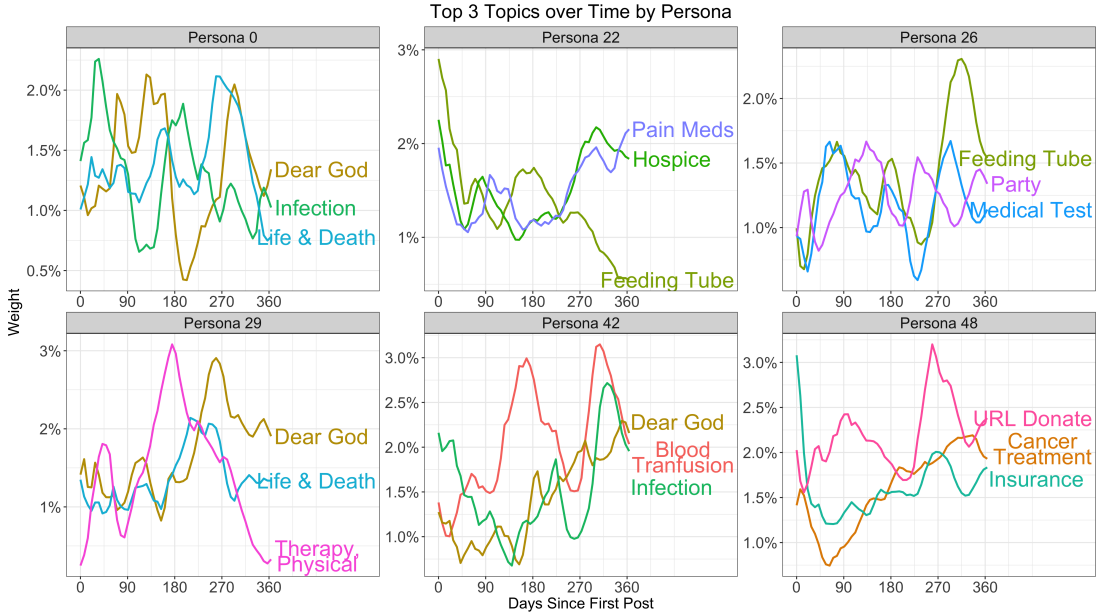


TABLE IV

COMPARISON OF DAPPER'S PERFORMANCE ON THE SM-BLOGS TEST CORPUS FOR VARYING NUMBER OF TOPICS, PERSONAS, AND BATCH SIZES. IN GENERAL, WE FIND THE FULL BATCH TRAINING ULTIMATELY LEADS TO HIGHER PER-WORD LOG-LIKELIHOODS. FOR ADDITIONAL HYPERPARAMETERS, PERSONAS HAS THE SMALLEST IMPACT ON PERFORMANCE, AND THE NUMBER OF TOPICS HAS NOTICABLE IMPACT. THE BEST THREE MODELS, IN TERMS OF HIGHEST TEST SET PWLL, ARE HIGHLIGHTED IN BOLD.

Number of Topics	Personas	Batch Size: 256	512	1024	2048	Full Batch
25	15	-6.46	-6.26	-6.17	-6.16	-5.65
25	25	-6.47	-6.26	-6.21	-6.23	-5.68
25	50	-6.55	-6.35	-6.34	-6.35	-5.71
50	15	-6.47	-6.30	-6.11	-6.16	**-4.97**
50	25	-6.50	-6.31	-6.30	-6.39	**-5.07**
50	50	-6.61	-6.38	-6.38	-6.52	-5.47
75	15	-6.87	-6.59	-6.46	-6.53	**-5.08**
75	25	-6.85	-6.61	-6.55	-6.61	-5.42
75	50	-6.96	-6.80	-6.79	-6.95	-6.00
100	15	-7.14	-6.93	-6.90	-6.98	-5.67
100	25	-7.07	-6.90	-6.91	-7.02	-5.99
100	50	-7.33	-7.17	-7.16	-7.31	-6.72

TABLE V

TOP EIGHT WORDS ASSOCIATED WITH THE MOST PREVALENT TOPICS FOUND BY THE DAPPER MODEL TRAINED ON THE FULL CARINGBRIDGE DATASET. TOPIC LABELS ARE SELECTED MANUALLY IN ORDER TO AID REFERENCE WITH FIGURE 3. THE WORDS _DOLLARS_, _NAME_, AND _URL_ REFER TO THE RESULT OF TEXT PRE-PROCESSING STEPS FOR CAPTURING COMMON PATTERNS LIKE THE DOLLAR AMOUNTS, ANONYMIZED NAMES, AND WEBSITE URLS, RESPECTIVELY.

Infection	Life & Death	Dear God	Pain Meds	Friend, Memories	Feeding Tube	Party	Medical Test
infection	life	god	cause	beautiful	tube	school	dr
fluid	live	lord	pain	friend	feed	birthday	test
lung	child	praise	medication	celebrate	breathe	fun	scan
remove	world	peace	brain	_name_	weight	_name_	result
procedure	others	pray	dose	card	oxygen	party	drug
chest	moment	trust	increase	memory	gain	aunt	mri
pressure	fear	father	level	flower	rate	game	ct
antibiotic	choose	joy	steroid	dance	ventilator	kid	liver

Therapy, Physical	Blood Tranfusion	Child	Hospice	Cancer Treatment	URL Donate	ICU	Insurance
therapy	blood	play	mom	cancer	_url_	icu	provide
physical	count	daddy	dad	radiation	_dollars_	brain	medical
leg	cell	mommy	visit	tumor	donate	monitor	information
therapist	low	girl	visitor	oncologist	money	stable	disease
arm	bone	boy	hospice	surgeon	benefit	wound	insurance
rehab	transplant	_name_	nursing	chemotherapy	en	neck	condition
foot	white	little	facility	breast	donation	doctor	regard
pt	marrow	cute	phone	biopsy	ha	unit	decision

sizes). Additionally, the DAPPER model requires time to be discretized in the data, and while the variational Kalman Filter keeps results from being too sensitive to the window size chosen, new methods exist capable of discretizing time based on shifts in topics [24].

ACKNOWLEDGMENTS

We thank reviewers for their valuable comments, University of Minnesota Supercomputing Institute (MSI) for technical support, and CaringBridge for their support and collaboration. The research was supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[2] M. D. Hoffman, D. M. Blei, and F. Bach, "Online Learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1–9, 2010.

[3] M. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic Variational Inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, 2012.

[4] J. D. Lafferty and D. M. Blei, "Correlated Topic Models," *Advances in Neural Information Processing Systems 18*, pp. 147–154, 2006.

[5] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," *International Conference on Machine Learning*, pp. 113–120, 2006.

[6] C. Wang, D. Blei, and D. Heckerman, "Continuous Time Dynamic Topic Models," *Proc of UAI*, pp. 579–586, 2008.

Fig. 4. Selected personas learned by the DAPPER model on the SM-blogs corpus. Each plot shows a different persona, and the three topics most strongly associated with that persona. For clarity, topic labels are hand-defined based on the top words in the topic and blog posts most associated with that topic (see Table VI). Results produce by DAPPER trained with full batch gradients, 50 topics, and 15 personas, which was found to perform best during compared to other hyperparameter settings (see Table IV for results on additional settings).

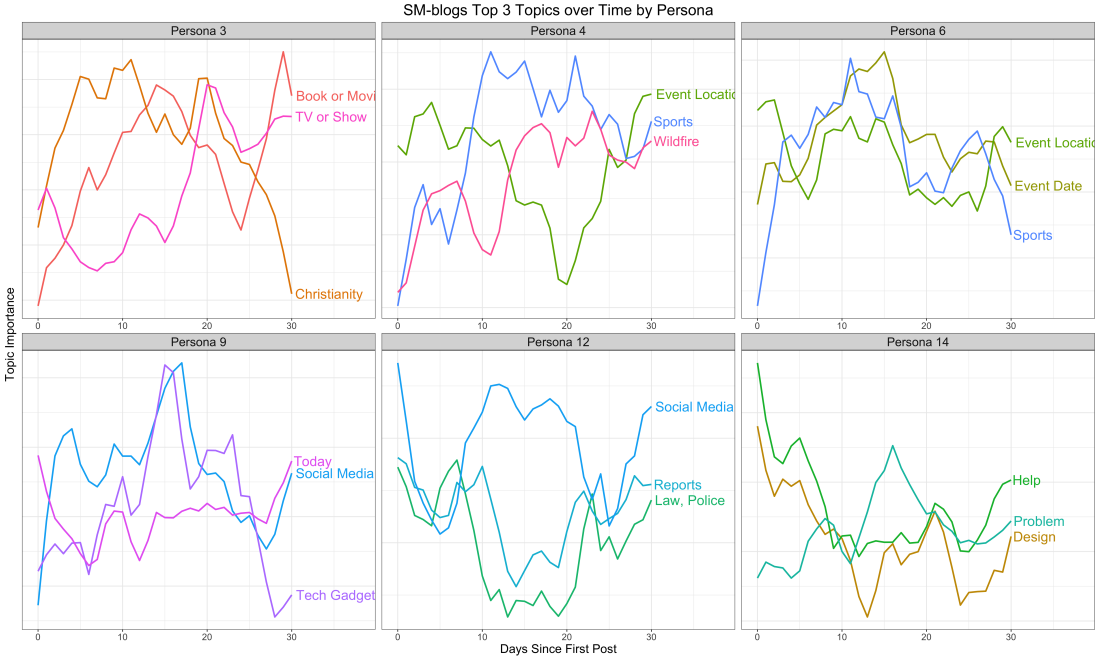


TABLE VI
TOP WORDS ASSOCIATED WITH THE MOST PREVALENT TOPICS FOUND BY THE DAPPER MODEL TRAINED ON THE FULL SM-BLOGS CORPUS. TOPIC LABELS ARE SELECTED MANUALLY IN ORDER TO AID REFERENCE WITH FIGURE 4.

Christianity	Tech Gadgets	Sports	TV or Show	Social Media	Wildfire	Book or Movie	Problem
life	apple	game	show	post	area	story	may
god	feature	season	live	share	water	book	change
word	phone	play	night	free	fire	movie	number
heart	device	team	star	comment	north	film	deal
church	user	against	news	video	land	full	result
pope	plus	player	series	photo	west	character	allow
father	update	football	special	click	near	author	note
christian	version	yard	tv	link	south	title	problem
son	iphone	coach	award	facebook	california	director	within
lord	app	ball	fan	twitter	local	writer	step

Law, Police	Today	Event Date	Design	Reports	Event Location	Help	Systems & Security
case	new	_year_	include	report	city	use	service
law	best	september	add	plan	event	need	system
police	today	th	design	issue	center	help	data
court	next	watch	create	member	st	different	technology
claim	open	online	large	public	street	small	customer
charge	month	date	image	accord	art	easy	network
officer	late	october	view	continue	park	save	access
against	hour	august	base	national	friday	important	solution
act	york	episode	space	action	sept	type	security
judge	check	july	form	official	monday	choose	provide

[7] R. Giaquinto and A. Banerjee, "Topic Modeling on Health Journals with Regularized Variational Inference," *AAAI*, 2018.

[8] M. E. Khan and W. Lin, "Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, pp. 878–887, 2017.

[9] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[11] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2007.

[12] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[13] R. Ranganath, S. Gerrish, and D. M. Blei, "Black Box Variational Inference," *Aistats*, vol. 33, 2013.

[14] M. E. Khan, R. Babanezhad, W. Lin, M. Schmidt, and M. Sugiyama, "Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Arlington, Virginia, United States: AUAI Press, jun 2016, pp. 319–328.

[15] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," vol. 2011, 2009.

[16] T. P. Minka, "A family of algorithms for approximate bayesian inference," *Ph.D. Thesis*, pp. 1–482, 2001.

[17] T. Minka, "Expecatation Propagation for Approximate Bayesian Inference," *Conference on Uncertainty in Artificial Intelligence*, pp. 362–369, 2001.

[18] M. Seeger, "Bayesian gaussian process models: Pac-bayesian generalisation error bounds and sparse approximations," 2003.

[19] A. Gelman, A. Vehtari, P. Jylänki, T. Sivula, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert, "Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data," 2014.

[20] A. Narayanan and V. Shmatikov, "Myths and fallacies of personally identifiable information," *Communications of the ACM*, vol. 53, no. 6, pp. 24–26, 2010.

[21] D. Corney, D. Albakour, M. Martinez, and S. Moussa, "What do a million news articles look like?" in *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, 2016, pp. 42–47. [Online]. Available: http://ceur-ws.org/Vol-1568/paper8.pdf

[22] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," *Advances in Neural Information Processing Systems 22*, pp. 288–296, 2009.

[23] S. Mandt and D. Blei, "Smoothed gradients for stochastic variational inference," in *Advances in Neural Information Processing Systems*, 2014, pp. 2438–2446.

[24] X. Chen, S. K. Candan, and L. M. Sapino, "Ims-dtm: Incremental multi-scale dynamic topic models." in *AAAI*, 2018.