

Detecting and Characterizing Events: Appendices

Anonymous EMNLP submission

A Inference

In this appendix, we describe the details of the variational inference algorithm for Capsule. This algorithm fits the parameters of the variational distribution q in Eq. 3 so that it is close in KL divergence to the posterior.

Recall that the variational distributions $q(\pi)$ and $q(\beta)$ are both Dirichlet-distributed with free variational parameters λ^π and λ^β , respectively. Similarly, the variational distributions $q(\psi)$, $q(\phi)$, $q(\theta)$ and $q(\epsilon)$ are all gamma-distributed with corresponding free variational parameters λ^ψ , λ^ϕ , λ^θ , and λ^ϵ . For these gamma-distributed variables, each free parameter λ has two components: shape s and rate r .

Minimizing the KL divergence between the true posterior p and the variational approximation q is equivalent to maximizing the ELBO (Eq. 2). We achieve this with closed form coordinate updates, as the Capsule model is specified with the required conjugate relationships that make this approach possible (Ghahramani and Beal, 2001).

To obtain simple updates, we first rely on auxiliary latent variables z . These variables, when marginalized out, leave the original model intact. The Poisson distribution has an additive property; specifically if $w \sim \text{Poisson}(a + b)$, then $w = z_1 + z_2$, where $z_1 \sim \text{Poisson}(z_1)$ and $z_2 \sim \text{Poisson}(z_2)$. We apply this decomposition to the word count distribution in Eq. 1 and define Poisson variables for each component of the word count:

$$z_{d,v,k}^{\mathcal{K}} \sim \text{Poisson}(\theta_{d,k} \beta_{k,v})$$

$$z_{d,v,t}^{\mathcal{T}} \sim \text{Poisson}(f(i_d, t) \epsilon_{d,t} \pi_{t,v}).$$

The \mathcal{K} and \mathcal{T} superscripts indicate the contributions from entity concerns and events, respectively. Given these variables, the total word count is deterministic:

$$w_{d,v} = \sum_{k=1}^K z_{d,v,k}^{\mathcal{K}} + \sum_{t=1}^T z_{d,v,t}^{\mathcal{T}}.$$

Coordinate-ascent variational inference is derived from complete conditionals, i.e., the conditional distributions of each variable given the other variables and observations. These conditionals define both the form of each variational factor and their updates. The following are the complete conditional for each of the gamma- and Dirichlet-distributed latent parameters. The notation $D(i)$ is used for the set of documents sent by entity i ; $D(t)$ is the set of documents sent impacted by events at time t (e.g., all documents after the event in the case of exponential decay).

$$\pi_t \mid \mathbf{W}, \psi, \phi, \beta, \theta, \epsilon, z \sim$$

$$\text{Dirichlet}_V \left(\alpha_\pi + \sum_{d=1}^D \langle z_{d,1,t}^{\mathcal{T}}, \dots, z_{d,V,t}^{\mathcal{T}} \rangle \right) \quad (4)$$

$$\beta_k \mid \mathbf{W}, \psi, \pi, \phi, \theta, \epsilon, z \sim$$

$$\text{Dirichlet}_V \left(\alpha_\beta + \sum_{d=1}^D \langle z_{d,1,k}^{\mathcal{K}}, \dots, z_{d,V,k}^{\mathcal{K}} \rangle \right) \quad (5)$$

$$\psi_t \mid \mathbf{W}, \pi, \phi, \beta, \theta, \epsilon, z \sim$$

$$\text{Gamma} \left(s_\psi + |D(t)| s_\epsilon, r_\psi + \sum_{d \in D(t)} \epsilon_{d,t} \right) \quad (6)$$

$$\phi_{i,k} \mid \mathbf{W}, \psi, \pi, \beta, \theta, \epsilon, z \sim$$

$$\text{Gamma} \left(s_\phi + |D(i)|s_\theta, r_\phi + \sum_{d \in D(i)} \theta_{d,k} \right) \quad (7)$$

$$\theta_{d,k} \mid \mathbf{W}, \psi, \pi, \phi, \beta, \epsilon, z \sim$$

$$\text{Gamma} \left(s_\theta + \sum_{v=1}^V z_{d,v,k}^{\mathcal{K}}, \phi_{a_d,k} + \sum_{v=1}^V \beta_{k,v} \right) \quad (8)$$

$$\epsilon_{d,t} \mid \mathbf{W}, \psi, \pi, \phi, \beta, \theta, z \sim$$

$$\text{Gamma} \left(s_\epsilon + \sum_{v=1}^V z_{d,v,t}^{\mathcal{T}}, \psi_t + f(i_d, t) \sum_{v=1}^V \pi_{t,v} \right) \quad (9)$$

The complete conditional for the auxiliary variables has the form $z_{d,v} \mid \psi, \pi, \phi, \beta, \theta, \epsilon \sim \text{Mult}(w_{d,v}, \omega_{d,v})$, where

$$\omega_{d,v} \propto \langle \theta_{d,1}\beta_{1,v}, \dots, \theta_{d,K}\beta_{K,v}, f(i_d, 1)\epsilon_{d,1}\pi_{1,v}, \dots, f(i_d, T)\epsilon_{d,T}\pi_{T,v} \rangle. \quad (10)$$

Intuitively, these variables allocate the data to one of the entity concerns or events, and thus can be used to explore the data.

Given these conditionals, the algorithm sets each parameter to the expected conditional parameter under the variational distribution. The mean field assumption guarantees that this expectation will not involve the parameter being updated. Algorithm 1 shows our variational inference algorithm.

References

Zoubin Ghahramani and Matthew J Beal. 2001. Propagation algorithms for variational bayesian learning. *Advances in neural information processing systems*, pages 507–513.

¹ $V(d)$ is the set of vocabulary indices for the collection of words in document d . We could also iterate over all V , but as zero word counts give $\mathbb{E}[z_{d,v}] = 0 \forall v \notin V(d)$, the two are equivalent.

Algorithm 1: Variational Inference for Capsule

Input: word counts w

Output: approximate posterior of latent parameters $(\psi, \pi, \phi, \beta, \theta, \epsilon)$ in terms of variational parameters $\lambda = \{\lambda^\psi, \lambda^\pi, \lambda^\phi, \lambda^\beta, \lambda^\theta, \lambda^\epsilon\}$

Initialize $\mathbb{E}[\beta]$ to slightly random around uniform

Initialize $\mathbb{E}[\psi], \mathbb{E}[\pi], \mathbb{E}[\psi], \mathbb{E}[\theta], \mathbb{E}[\epsilon]$ to uniform

for iteration $m = 1 : M$ **do**

set $\lambda^\psi, \lambda^\pi, \lambda^\phi, \lambda^\beta, \lambda^\theta, \lambda^\epsilon$ to respective priors, excluding $\lambda^{\theta, rate}$ and $\lambda^{\epsilon, rate}$, which are set to 0

update $\lambda^{\theta, rate} += \sum_V \mathbb{E}[\beta_v]$

for each document $d = 1 : D$ **do**

for each term $v \in V(d)$ ¹ **do**

set $(K + T)$ -vector $\omega_{d,v}$ using $\mathbb{E}[\pi]$, $\mathbb{E}[\theta]$, and $\mathbb{E}[\epsilon]$, as shown in Eq. 10

set $(K + T)$ -vector

$\mathbb{E}[z_{d,v}] = w_{d,v} * \omega_{d,v}$

update $\lambda_d^{\theta, shape} += \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 8)

update $\lambda_d^{\epsilon, shape} += \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ (Eq. 9)

update $\lambda_v^\beta += \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 5)

update $\lambda_v^\pi += \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ (Eq. 4)

end

update $\lambda_d^{\theta, rate} += \mathbb{E}[\phi_{a_d}]$ (Eq. 8)

update $\lambda_d^{\epsilon, rate} += \mathbb{E}[\psi]$ (Eq. 9)

set $\mathbb{E}[\theta_d] = \lambda_d^{\theta, shape} / \lambda_d^{\theta, rate}$

set $\mathbb{E}[\epsilon_d] = \lambda_d^{\epsilon, shape} / \lambda_d^{\epsilon, rate}$

update $\lambda_{a_d}^{\phi, shape} += s_\theta$ (Eq. 7)

update $\lambda_t^{\psi, shape} += s_\epsilon \forall t : f(i_d, t) \neq 0$ (Eq. 6)

update $\lambda_{a_d}^{\phi, rate} += \theta_d$ (Eq. 7)

update $\lambda^{\psi, rate} += \epsilon_d$ (Eq. 6)

end

set $\mathbb{E}[\phi] = \lambda^{\phi, shape} / \lambda^{\phi, rate}$

set $\mathbb{E}[\beta_k] = \lambda^{\beta_k, v} / \sum_v \lambda^{\beta_k} \forall k$

set $\mathbb{E}[\psi] = \lambda^{\psi, shape} / \lambda^{\psi, rate}$

set $\mathbb{E}[\pi_t] = \lambda^{\pi_t, v} / \sum_v \lambda^{\pi_t} \forall t$

end

return λ
