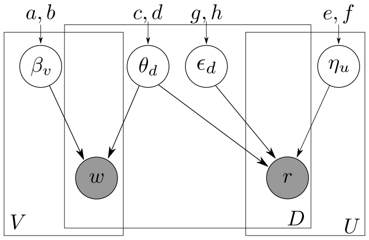

Figure 1: The collaborative topic Poisson factorization model (CTPF).

A Stochastic variational inference for the collaborative topic Poisson factorization model

Stochastic variational inference combines stochastic gradient algorithms and variational inference [3]. Stochastic gradient algorithms follow noisy estimates of the gradient with a decreasing step-size. If the expectation of the noisy gradient equals to the gradient and if the step-size decreases according to a certain schedule, then the algorithm converges to a local optimum [4]. To obtain noisy gradients, assume that we operate under the setting where we subsample a single document d uniformly at random from the D documents. This sampling strategy is similar to online LDA [2]. However, our approach differs in the use of separate learning rates for each user, allowing the inference to update only the relevant users in each iteration.

We are given observations about a single document in each iteration. Following [3], we use the conditional dependencies in our graphical model to divide our variational parameters into *local* and *global*. The multinomial parameters (ϕ_{dv}, ξ_{ud}) for the sampled document d and for all $u \in U$, and the Gamma parameters for $(\theta_{dk}, \epsilon_{dk})$ are local. All other varational parameters are global.

In each iteration of our algorithm, we first subsample a document. We then update the local multinomial parameters and the local topic intensities and offset parameters for this document using the coordinate updates from Figure 2. This optimizes the local parameters with respect to the subsample. We then compute scaled natural gradients [1] for the global user preference parameters $(\tilde{\eta}^{\text{shp}}_{uk}, \tilde{\eta}^{\text{rte}}_{uk})$ for the users u that have rated document d and for all topic parameters $(\tilde{\beta}^{\text{shp}}_{vk}, \tilde{\beta}^{\text{rte}}_{vk})$. The global step for the global parameters follows the noisy gradient with an appropriate step-size.

We maintain separate learning rates ρ_u for each user, and only update the users who have rated the document d. We proceed similarly for words. We maintain a global learning rate ρ' for the topic parameters, which are updated in each iteration. For each of these learning rates ρ, we require that $\sum_t \rho(t)^2 < \infty$ and $\sum_t \rho(t) = \infty$ for convergence to a local optimum [4]. We set $\rho(t) = (\tau_0 + t)^{-\kappa}$, where $\kappa \in (0.5, 1]$ is the learning rate and $\tau_0 \geq 0$ downweights early iterations [3].

References

[1] S. Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 10(2):357–385, 1982.

[2] M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2010.

[3] M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[4] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.