Online Inference for Relation Extraction with a Reduced Feature Set

Maxim Rabinovich
Computer Science Division
University of California, Berkeley
rabinovich@eecs.berkeley.edu

Cédric Archambeau*
Amazon
Berlin, Germany
cedrica@amazon.de

Abstract

Access to web-scale corpora is gradually bringing robust automatic knowledge base creation and extension within reach. To exploit these large unannotated—and extremely difficult to annotate—corpora, unsupervised machine learning methods are required. Probabilistic models of text have recently found some success as such a tool, but scalability remains an obstacle in their application, with standard approaches relying on sampling schemes that are known to be difficult to scale. In this report, we therefore present an empirical assessment of the sublinear time sparse stochastic variational inference (SSVI) scheme applied to RelLDA. We demonstrate that online inference leads to relatively strong qualitative results but also identify some of its pathologies—and those of the model—which will need to be overcome if SSVI is to be used for large-scale relation extraction.

1 Introduction

Access to web-scale corpora is gradually bringing automatic knowledge base creation and extension within reach (Mausam et al., 2012). Human curated resources, such as Freebase (Bollacker et al., 2008), are invaluable for relation extraction, but they are inherently incomplete. The total number of relations that might be encountered is unbounded and the number actually encountered in a corpus grows with its size. Hence the need for unsupervised methods and the recent small-scale success on this problem with probabilistic models. Unfortunately, prohibitive memory usage and training time makes their large-scale application all but impossible, and the incremental training algorithms used to train topic models like latent Dirichlet allocation (LDA) at scale (Hoffman et al., 2010; Mimno et al., 2012) have not yet been applied to relation extraction.

In this paper, we show that sparse stochastic variational inference (SSVI) (Mimno et al., 2012) can be applied to the RelLDA model for unsupervised relation extraction introduced by (Yao et al., 2011; Yao et al., 2012). SSVI is attractive for two reasons. First, it processes corpora incrementally, speeding convergence and supporting streaming. Second, it improves on plain stochastic variational inference by using sparse updates able to deal with a large number of topics. We find that our algorithm is able to obtain strong qualitative results in a fraction of the time that is needed to run the Gibbs sampler for RelLDA and with a reduced memory footprint. We also include discussion of some pitfalls in unsupervised relation extraction with LDA-style models and how they might be overcome, and we show that dependency parse features are not needed for this task, a major departure from prior work in this area.

2 Model Specification

We use a modified form of RelLDA (Yao et al., 2011), eliminating the reliance on a dependency parsed corpus. Relations are grouped into clusters. Each document is assumed to behave as a mixture of these relation clusters, with each sentence in the document exhibiting exactly one of them. Multiple feature sets are permitted, which we exploit below to use separate vocabularies for entity features, linking word features, and syntactic features. Throughout this paper, we adopt the convention that R refers to the number of relation clusters, F to the number of feature types, W_f to the vocabulary size for feature type f ($1 \leq f \leq F$), N_d to the number of sentences in a document,

* Work undertaken while the second author was at Xerox Research Centre Europe, supervising the first author's research internship.

and N_{dif} to the number of features of type f exhibited by sentence i in document d.

In this notation, the relation clusters are defined as a set of F discrete distributions over the feature vocabularies:

For $r = 1, \ldots, R$ and $f = 1, \ldots, F$:
Draw $\beta_{rf} \sim \text{Dirichlet}(\eta_f)$,

where $\eta_f > 0$ is a scalar.[1] The generative process for relations takes the following form:

For $d = 1, \ldots, D$:

1. Draw $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For $i = 1, \ldots, N_d$:
   (a) Draw $z_{di} \sim \theta_d$.
   (b) For $f = 1, \ldots, F$ and $j = 1, \ldots, N_{dif}$: Draw $w_{dij} \sim \beta_{z_{di}f}$.

where $\alpha > 0$ is again a scalar, and θ_d defines a discrete distribution over the relation clusters associated to document d. The relation in each sentence is drawn from θ_d and the associated features from β_{rf}.

2.1 Extracting entity pairs

We assume access to a part-of-speech (POS) tagger and a named entity recognizer (NER). Our ultimate goal is to extract relations between named entities and therefore necessarily limit attention to sentences with at least two entity mentions. Sentences with more than two mentions pose a problem due to *a priori* ambiguity in the pairs being related, so we simply assume the salient entity pair is the one that is closest together in the sentence—a simple heuristic that allows us to avoid modeling sentence segmentation. We use the Stanford CoreNLP library for both POS tagging and NER (Finkel et al., 2005).

2.2 Feature sets

Our experiments all draw on feature types built according to a small set of templates and always reflecting only the sequence of words *between* two selected entity mentions in the sentence:

Entity surface strings. Each sentence contains two distinguished entity mentions. The left (first) and right (second) strings are treated as features of distinct types to capture asymmetry. The vocabularies for those two types are, however, the same. We refer to the resultant features as ENT^{left} and ENT^{right} .

Entity types. The Stanford NER outputs entity types in { PER , ORG , LOC , MISC }, referring to the **per**son, **org**anization, **loc**ation, and **misc**ellaneous, respectively. We use the pair $(t_1,\ t_2)$ of entity types for the two distinguished entities as a feature. This feature type is referred to as ENT-TYPE .

Phrases between the entities. The word sequence between the entities is partitioned into coarse-grained part-of-speech categories: ADJ (JJ, JJR, JJS), ADV (RB, RBR, RBS), NN (NN, NNS, NNP, NNPS, PRP, WP), PP (IN, TO), VB (VB, VBD, VBG, VBN, VBP, VBZ), and OTH (everything else). We refer to the resultant six feature sets as ADJ , ADV , NN , OTH , PP , and VB .

POS tag sequences. We include a feature corresponding to the entire sequence of Penn Treebank POS tags between the two entities. We refer to this feature type as POS-SEQ .

3 Sparse Stochastic Variational Inference

To make inference scalable to very large corpora, we use the sparse stochastic variational inference (SSVI) originally developed for LDA (Mimno et al., 2012). The true posterior over $\beta_{1:R,1:F}$ is approximated by a product of independent Dirichlets, viz.

$$q(\beta_{1:R,1:F}) = \prod_{r=1}^{R} \prod_{f=1}^{F} q(\beta_{rf}),$$

where $q(\beta_{rf}) = \text{Dirichlet}(\lambda_{rf})$ and $\lambda_{rf} \in \mathbb{R}_+^{W_f}$ are variational parameters. Classical variational Bayes would also approximate the posterior over θ_d and z_{di} by Dirichlet and multinomial distributions, respectively, leading to $\Omega(DR)$ memory usage and $\Omega(R)$ time for local updates. SSVI reduces both requirements to $O(1)$ by eliminating the local variational distribution. Instead, it integrates out θ_d and uses samples from the an optimized variational distribution $q^*(z_d)$ to estimate the expectations required in the updates. Here the optimality criterion for q^* is simply that its Kullback-Leibler divergence from the true z_d posterior is as small as possible within the constraints imposed by its factored form (Bishop, 2006).

[1] Note that this means we are using a symmetric Dirichlet, viz. $p(\beta \mid \eta) \propto \prod_v \beta_v^{\eta-1}$.

Furthermore, the entire corpus need not be considered during each step; rather, a random minibatch $B = \{d_1, \ldots, d_S\}$ of documents is considered and sampling is carried out only for those documents. Each iteration thus only needs to update the parameters associated with relations r, features types f, and features values v encountered in B. This leads to the following variational updates:

$$\lambda_{rfv}^{(t+1)} = (1 - \rho^{(t)})\lambda_{rfv}^{(t)} + \rho^{(t)} \cdot \frac{D}{S} \sum_{d \in B} \hat{\mathbb{E}}\left[N_{drfv}\right],$$

where $\rho^{(t)}$ is the learning rate, N_{drfv} is the number of times feature value v of type f is assigned to relation r in document d and $\hat{\mathbb{E}}$ denotes a Monte Carlo estimate of an expectation. Using a trick we explain in the supplement, we can ensure that each iteration only updates parameters λ_{rfv} for relations r, feature types f, and feature values v that occur in that iteration's minibatch (the origin of the *sparse* moniker). The supplement likewise explains our natural gradient hyperparameter optimization scheme for η_f and α.

4 Empirical Evaluation

4.1 Datasets

We use the AQUAINT2 2 corpus, consisting of articles from several newspapers including the New York Times (Vorhees and Graff, 2008). After eliminating sentences with fewer than two entities, we were left with 578790 documents (1492599 sentences), of which 462755 (1193275 sentences) were used in training and the remainder used for evaluation. The sizes of the feature sets for this data were: 8996 (ADJ), 7334 (ADV), 233725 (ENT^{left}), 233725 (ENT^{right}), 39895 (NN), 52998 (OTH), 16564 (PP), 28826 (VB), 89022 (POS-SEQ), and 16 (ENT-TYPE). We consider two subset of the features in our experiments:

1. The full feature set: ADJ , ADV , ENT^{left} , ENT^{right} , OTH , PP , VB , POS-SEQ , and ENT-TYPE .

2. All features excluding the entity features: ADJ , ADV , OTH , PP , VB , POS-SEQ , and ENT-TYPE .

4.2 Model selection

The hyperparameters are optimized as part of the algorithm. SSVI includes a learning rate $\rho^{(t)}$ generally set to

$$\rho^{(t)} = \frac{a}{(b + t)^c},$$

where $a, b > 0$ and $\frac{1}{2} < c \leq 1$. This choice of schedules allows convergence of the algorithm to a local optimum of the objective to be guaranteed (Hoffman et al., 2013). In practice, setting c at or close to $\frac{1}{2}$ give good results.

We fit the model with several values of R, a, and b and score each based on its perplexity and variational objective values on an evaluation corpus. We carried out a grid search for values with $R \in \{250, 500, 1000\}$, $a \in \{0.1, 0.01, 0.001\}$, and $b \in \{1.0, 10.0\}$. We find that the choice of these parameters has a noticeable but not substantial effect on the metrics. Nonetheless, we limited our qualitative evaluation to the best learning rates in terms of the variational objective (-9.02×10^6), that is, $a = 0.01, b = 10.0$ and $K = 500$. The number of iterations T of SSVI, on the other hand, had a substantial effect. Figure 1 illustrates this with varying values of R and $a = 0.1,\ b = 1.0$.

4.3 Discovered relations

Evaluating the quality of the relations discovered by our algorithm is challenging in the absence of ground truth, especially due to the inherent noisiness of relation clusters discovered by any unsupervised learning algorithm—and by stochastic gradient methods in particular. Ordinarily, the output of LDA-type models is shown as per-topic rankings of the vocabulary. In our setting, this makes little sense due to the multi-view setup and the fact that, e.g., the most likely entities under a relation need not correspond to the most likely noun phrases. We thus represent relation clusters as lists of sentences most strongly associated with them. The strength of association was determined by taking 50 posterior samples of the relation assignment for each sentence and computing the proportion of samples assigned to each relation.

As Table 1 shows, the clusters are reasonably coherent but quite noisy. The first corresponds to a general constellation of relations between people and organizations that could reasonably be summarized as "occupies leadership position at," though in reality, the generalization made by the inference procedure is somewhat narrower than that, with a bias toward political leaders. The second is much more restricted and basically corresponds to the concept of being a "market strategist

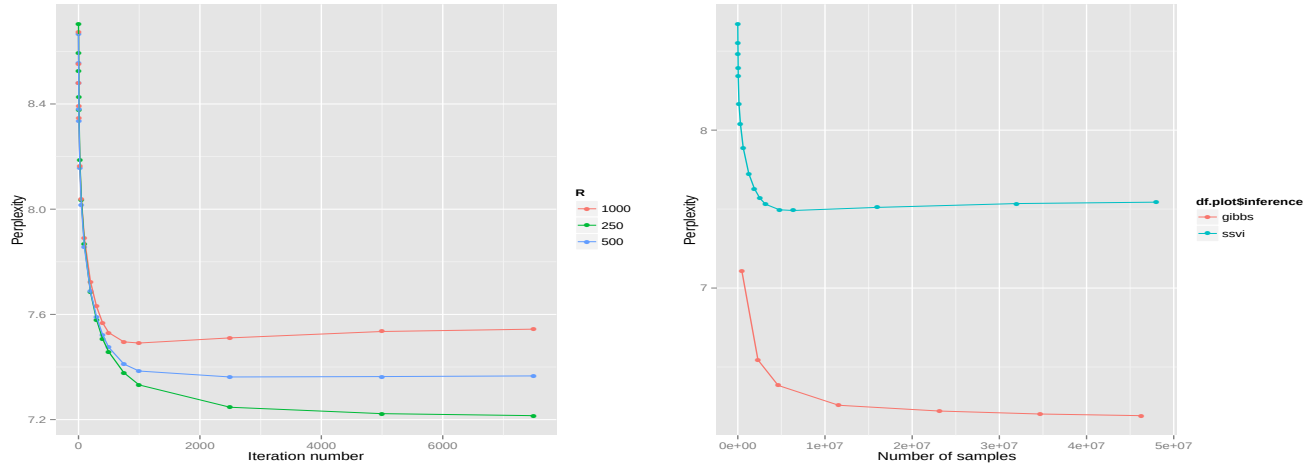

Figure 1: *(Left)* Perplexity on an evaluation corpus for SSVI as a function of iteration ($a = 0.01,\ b = 1.0$). *(Right)* A comparison of evaluation perplexity for SSVI and Gibbs sampling with $R = 1000$.

`leader-at` relation
European / Peter Mandelson / trade commissioner / NN NN OT / MISC-PER
UN / Joao Bernardo Honwana / special envoy / NN NN / ORG-PER
UN / Pierre Goldschmidt / director general / 's deputy / ORG-PER
Zimbabwe / Morgan Tsvangirai / opposition leader / NN NN / LOC-ORG
Spanish / Jose Antonio Alonso / counterpart , / NN OT / MISC-PER
European Union / Pascal Lamy / trade commissioner / NN NN / ORG-PER
ASIO / Dennis Richardson / director general / OT NN JJ / ORG-PER
WTO / EU / trade commissioner / NN NN OT / MISC-PER
UN / Jacques Klein / special envoy / NN NN / ORG-PER
pro-Russian / Viktor Yanukovich / opposition leader / NN NN / MISC-PER

`trader-at/market-strategist-at` relation
Roma / Livorno / bottom club / 2-0 away to / VBD CD RB TO NN NN / ORG-LOC
Art Hogan / Jefferies and Co. / market strategist at / , chief / OT JJ NN NN IN / PER-ORG
Kenneth Tower / CyberTrader / market strategist at / , chief / OT JJ NN NN IN / PER-ORG
Chinese / Ssangyong / bidder for / firm , / as the / NN OT VBD IN DT JJ NN IN / MISC-ORG
Michael Sheldon / Spencer Clark LLC / market strategist at / , chief / OT JJ NN NN IN / PER-ORG
Oracle Corp. / PeopleSoft / business software maker / bid for / ORG-ORG
US / Asian / market strategist at / , chief / OT JJ NN NN IN / PER-ORG
SAIC / Birmingham-based / automaker , / fortunes of / one billion / that could potentially / 1.85 billion / ORG-MISC
Barry Ritholtz / Maxim Group / market strategist at / , chief / OT JJ NN NN IN / PER-ORG
Al Goldman / AG Edwards / market strategist at / , chief / OT JJ NN NN IN / PER-ORG

Table 1: Sentences in the corpus most strongly associated with one of the relations, as determined by sampling relation assignments. *(Top)* This particular relation appears to identify the concept of "occupies leadership position at," while the second relation *(bottom)* appears to identify the concept of "trader at" or "trading strategist at." Parameters were set to $R = 500$, $a = 0.009$, and $b = 10.0$ for both.

at." Even so, the model picks up on the fact that "bidder for" is a closely related concept and expresses a similar relationship between the person and organization in question.

4.4 Clustering pathologies

The results we show correspond to a feature set excluding ENT^{left} and ENT^{right} , as we found certain pathologies in the output with the full feature set, notably a tendency for some relation clusters to form around sets of entities rather than the relations between them. Figure 2 illustrates this effect.

Removing entity features resolves this first issue. Overcoarsening of relation clusters is a more persistent problem. Some relations look more like broad topics than focused relations. The likely cause of this is allocation of topic words that co-occur with relation words to that relation due to the absence of a special set of shared topic distributions that could catch the intruding words. Figure 2 illustrates this problem within one relation.

The incorporation of syntactic features is another source of over-coarsening, with some relation clusters forming around common syntactic patterns. The best illustration of this are the POS tag sequences "NNS IN" and "NN IN", which served as the basis for clustering of unrelated concepts like "headquarters in," "crisis in," and "meeting in."

At their core, the pathologies we uncover all appear to flow from problems in the model rather than the inference scheme—notably the requirement that each word be explained by a relation. The absence of any broader shared topic distributions that can be used to explain away non-relation-specific words causes some relations to behave very much like topics and all relations to catch many co-occurrent words that are not essentially part of their semantic content. Likewise, although the addition of syntactic features allows abstraction away from specific word patterns to more generally applicable syntactic ones, it also

Entity-based relation	Topic-like relation
French Riviera / Cannes / resort of Northern Gaza / Israeli / withdrawal of the / and the Gaza / Israeli / withdrawal of the / and the France / China / deficit with	policemen and / city of /near the / were killed blows himself / suicide bomber / up during / when a incursion into / the northern rebel stronghold of / roadside bomb / were wounded

Figure 2: *(Left)* A relation based on sets of related entities. *(Right)* A topic-like relation. Both exclude POS-SEQ and ENT-TYPE features for brevity.

leads to problems if the model does not account for syntactic overlap of semantically distinct relations. Both of these issues could be addressed by adding additional hierarchy to the model. For instance, a set of global topic distributions could be added to resolve the first problem, while relations could be grouped into higher-level clusters governing syntactic properties to resolve the other. We believe such modifications are likely to lead to much more robust models of relations in text without significantly complicating inference.

4.5 Comparison to Gibbs sampling

Since we use a minibatch size of $S = 256$ and $S' = 25$ samples to form our estimate of the natural gradient, each iteration of SSVI corresponds to 6400 document steps in the Gibbs chain. As a result, one full Gibbs sweep through the corpus is equivalent to about 75 SSVI iterations in terms of numbers of samples taken.[2] We use this as the basis of the plot in Figure 1.

Surprisingly, Gibbs sampling appears to achieve better held-out perplexity at each given level of computation. This is contrary to the expected behavior of SSVI (Hoffman et al., 2010; Mimno et al., 2012) and does not have a clear explanation. The most likely causes are, first, the learning parameters, as stochastic gradient methods are known to be extremely sensitive to the choice of learning rate (Ranganath et al., 2013) and, second, the inherent noisiness of stochastic gradient methods, which work best on large, highly redundant corpora. Although we lightly optimized the learning parameters, it is possible that more extensive experiments would discover a drastically better setting of those parameters; alternatively, adaptive rate methods may be needed. If, on the other hand, is simply the noisiness of the stochastic gradients, then variance reduction techniques may yield better results (Paisley et al., 2012).

It is also important to account for the computational aspect of the performance metric. Often, one can drastically reduce the size of the minibatches in SSVI (e.g. to 64 documents), which would lead to multiplicative speedups (e.g. 4x if $S = 64$); likewise, the number of Gibbs sweeps used for estimation on the minibatch could be reduced, as could the burnin for those sweeps. Such fine-tuning is beyond the scope of this work but, based on our results, would be a crucial component in practical systems seeking to reap the benefits of SSVI with models like RelLDA.

Finally, it may simply be that for a complex model like RelLDA, the data set must be made far larger before sufficient redundancy appears, in which case we would expect to see gains in the relative performance of SSVI and Gibbs sampling in the regime of larger information extraction datasets, which often contain hundreds of millions of documents. This last point also illustrates how SSVI might be advantageous even if less statistically efficient: unlike Gibbs sampling, whose memory usage grows with the size of the corpus, SSVI can operate with a fixed amount of memory—just enough to store the minibatch data structures.

5 Conclusion

We have shown that SSVI is a promising technique for relation extraction at scale. Apart from some pathologies due to the modeling assumptions, it discovers coherent relational clusters while requiring less memory and time than sampling methods. Moreover, the issues we uncover point to problems with the model that suggest how more effective probabilistic models of relations in text might be designed and used.

[2] A more exact number is $\frac{462755}{6400} \approx 72.3$.

Acknowledgements

The author would like to thank Guillaume Bouchard for helpful discussions about improvements to the core model.

Appendices

In the following appendices, we explain the mathematics of our inference algorithm in detail.

A The core algoritihm

An alternative to MAP inference via Gibbs sampling is variational inference. Ordinarily, this would be done by specifying a variational distribution over the β, z, and θ variables. In our setup, because each observation consists of multiple words, the standard way of doing this fails, however. Fortunately, we can use a recent stochastic approach that still works and that scales much better than batch variational Bayes. This approach is based on the strategy for LDA set out in (Mimno et al., 2012).

To do this, we posit

$$q(\beta_{rf}) = \mathrm{Dir}(\lambda_{rf}), \;\; \lambda_{rf} \in \mathbb{R}_+^{V_f}$$

and let $q(z_d)$ be an arbitrary distribution, which will be chosen to be the optimal one per the analytical (but uncomputable) variational Bayes update formula. The mixing distributions θ are marginalized out as in collapsed Gibbs sampling. Since our goal is to optimize λ, we write the ELBO up to a constant independent of λ:

$$\mathcal{L} = \sum_d \sum_{r,f} \left[\sum_v \left(\mathbb{E}\left[N_{drfv}\right] + \frac{\eta_f - \lambda_{rf}}{D} \right) \times \mathbb{E}_q\left[\log \beta_{rfv}\right] + \frac{1}{D}\left(\sum_v \log\Gamma(\lambda_{rfv}) - \log\Gamma(\Lambda_{rf}) \right) \right],$$

where $\Lambda_{rf} = \sum_v \lambda_{rfv}$. We know $\mathbb{E}_q[\log \beta_{rfv}] = \Psi(\lambda_{rfv}) - \Psi(\Lambda_{rf})$, and we use sampling over z_d to approximate $\mathbb{E}_q[N_{drfv}]$. Specifically, basic theory tells us that the optimal choice of variational distribution over z_d, holding those over all other latent variables fixed, is

$$\begin{aligned} q^*(z_d) &\propto \exp\left(\mathbb{E}_{q^{\setminus z_d}}[\log p(z_{1:D},\ v_{1:D},\ \beta_{1:R,\ 1:F}]\right) \\ &\propto \exp\left(\mathbb{E}_{q^{\setminus z_d}}\left[\log p(v_d \mid z_d,\ \beta) + \log p(z_d \mid \alpha)\right]\right) \\ &\propto p(z_d \mid \alpha) \prod_{r,\ f} \prod_{v:\ N_{drfv}>0} \exp\left(N_{drfv} \mathbb{E}_q\left[\log \beta_{rfv}\right]\right) \\ &= \left(\frac{\Gamma(R\alpha)}{\Gamma(O_d + R\alpha)} \cdot \prod_r \frac{\Gamma(O_{dr} + \alpha)}{\Gamma(\alpha)} \right) \\ &\quad \times \prod_{r,\ f} \prod_{v:\ N_{drfv}>0} \exp\left(N_{drfv}\left[\Psi(\lambda_{rfv}) - \Psi(\Lambda_{rf})\right]\right). \end{aligned}$$

We thus find

$$q^*(z_{do} = r \mid z_d^{\setminus do}) \propto (O_{dr} + \alpha) \times \prod_f \prod_{v:\ N_{dofv}>0} \exp\left(N_{dofv}\left[\Psi(\lambda_{rfv}) - \Psi(\Lambda_{rf})\right]\right), \tag{1}$$

which means we can approximately sample from q^* using Gibbs sampling to obtain an approximation to $\mathbb{E}_q[N_{drfv}]$.

Why is this helpful? As shown in (Hoffman et al., 2013), the natural gradient of $\mathcal{L}$ in the rfv dimension is given by

$$\mathbb{E}_{q(z_{1:D})}\left[\sum_d N_{drfv}\right] + \eta - \lambda_{rfv}.$$

Split up over documents, this gives a per-document contribution of

$$\mathbb{E}_{q(z_d)}\left[N_{drfv}\right] + \frac{1}{D}\left(\eta - \lambda_{rfv}\right).$$

This means that if we sample a batch of documents $d_1, \ldots, d_S$ and approximate $\mathbb{E}_{q(z_d)}\left[N_{drfv}\right]$ using S' rounds of Gibbs sampling, we will end up with an unbiased estimate of the natural gradient that we can use for stochastic gradient ascent on $\mathcal{L}$. With a little bit more work, we can make all necessary updates sparse to ensure efficiency.

Concretely, each iteration of the algorithm does the following.

1. Sample a minibatch $\mathcal{M} = \{d_1, \ldots, d_S\}$ of S documents (without replacement).

2. Run B burn-in rounds of Gibbs sampling on z_d for $d \in \mathcal{M}$ using (1). Then run S' more sweeps, saving the value of $\sum_{d\in\mathcal{M}} N_{drfv}$ after each one. Estimate $\sum_{d\in\mathcal{M}} \mathbb{E}_q[N_{drfv}]$ by $\hat{N}^{\mathcal{M}}_{drfv} :\ = \frac{1}{S'}\sum_{s'=1}^{S'} N^{(s')}_{drfv}$.

3. Estimate the rfv component of the overall natural gradient by

$$\hat{g}_{rfv} \colon = \frac{D}{S} \cdot \hat{N}^{\mathcal{M}}_{rfv} + \eta_f - \lambda_{rfv}.$$

4. Update

$$\lambda_{rfv} \leftarrow \lambda_{rfv} + \rho \hat{g}_{rfv},$$

where $\rho = \rho_t$ is the current learning rate.

Note that if we write $\hat{N}_{drfv} = \frac{D}{S} \cdot \hat{N}^{\mathcal{M}}_{rfv}$ and let $\tilde{N}_{rfv} = \lambda_{rfv} - \eta_f$ (this is the pseudocount part of the variational parameter), we have $\lambda_{rfv} = \tilde{N}_{rfv} + \eta_f$ and hence an update of the form

$$\tilde{N}_{rfv} \leftarrow (1-\rho)\tilde{N}_{rfv} + \rho \hat{N}_{rfv}.$$

Note further that if we let $\pi_t = \prod_{\tau=0}^{t} (1-\rho_\tau)$, we can write this update as

$$\frac{\tilde{N}^{(t)}_{rfv}}{\pi_t} = \frac{\tilde{N}^{(t-1)}_{rfv}}{\pi_{t-1}} + \frac{\rho \hat{N}^{(t)}_{rfv}}{\pi_t}.$$

Thus, if we track $\frac{\tilde{N}^{(t)}_{rfv}}{\pi_t}$ rather than the raw pseudocount, we get sparse updates. This is what the code actually does.

B Adding hyperparameter optimization

In its current form, the variational inference algorithm requires the Dirichlet hyperparameters η_f to the global relation distributions β_{rf} and α to the local mixing distributions θ_d to be set manually. To remove this limitation, we extend the natural gradient descent scheme to the hyperparameters.

To begin, note that the part of the variational objective that depends on η_f is given by

$$\mathcal{L}(\eta_f) = \eta_f \cdot \sum_r \sum_v \left[\Psi(\lambda_{rfv}) - \Psi(\Lambda_{rf})\right] - R \cdot \left[V_f \cdot \log\Gamma(\eta_f) - \log\Gamma(V_f\eta_f)\right],$$

whence

$$\frac{\partial \mathcal{L}}{\partial \eta_f} = \sum_r \left[\sum_v \left[\Psi(\lambda_{rfv}) - \Psi(\eta_f)\right] - V_f \cdot \left[\Psi(\Lambda_{rf}) - \Psi(V_f\eta_f)\right] \right].$$

However, we would like to use natural gradient updates, which have the form

$$\eta_f^{(t+1)} = \eta_f^{(t)} + \rho_t \left[G^{(t)}_{\eta,f}\right]^{-1} \nabla_{\eta_f} \mathcal{L},$$

where $G^{(t)}_{\eta,f} = \mathbb{E}\left[\left(\frac{\partial \log p(\beta_f \mid \eta_f)}{\partial \eta_f}\right)^2 \mid \eta_f\right](\eta_f^{(t)})$ is the Fisher information matrix for the parameter η_f evaluated at the value $\eta_f^{(t)}$. Since

$$\log p(\beta_f \mid \eta_f) = (\eta_f - 1) \cdot \sum_{r,v} \log \beta_{rfv} - R\left(V_f \log\Gamma(\eta_f) - \log\Gamma(V_f\eta_f)\right).$$

This is easy to compute.

Indeed, if we write $\log p(\beta_f \mid \eta_f) = t(\beta_f)\cdot\eta_f - t(\beta_f) - a(\eta_f)$ with $t(\beta_f) = \sum_r \sum_v \log \beta_{rfv}$, we need only compute $\mathbb{E}\left[\left(t(\beta_f) - a'(\eta_f)\right)^2\right]$, which, by the usual exponential family identities, is given by

$$\mathbb{E}\left[t(\beta_f)^2\right] - \mathbb{E}\left[t(\beta_f)\right]^2 = a''(\eta_f).$$

Fortunately, we know

$$a'(\eta_f) = RV_f \cdot \left[\Psi(\eta_f) - \Psi(V_f\eta_f)\right],$$

so we can calculate

$$G_{\eta,f}(\eta_f) = a''(\eta_f) = RV_f \cdot \left[\psi_1(\eta_f) - V_f\psi_1(V_f\eta_f)\right],$$

where $\psi_1 = \Psi'$ is the first polygamma function (the trigamma function). Note that, analogously,

$$G_\alpha(\alpha) = DR \cdot \left[\psi_1(\alpha) - R\psi_1(R\alpha)\right].$$

The (unnatural) gradient for α is harder to compute, however:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_d \frac{\partial}{\partial \alpha} \mathbb{E}_q\left[\log p(z_d \mid \alpha)\right] = \sum_d \mathbb{E}_q\left[\frac{\partial}{\partial \alpha} \log p(z_d \mid \alpha)\right].$$

Since the expectation cannot be analytically computed, we use our samples $z_d^{(s')}$ for $s' = 1, \ldots, S'$ and $d \in \mathcal{M}$ to compute a stochastic gradient. For this, we first note that for fixed z_d,

$$\log p(z_d \mid \alpha) = \sum_r \left[\log\Gamma(O_{dr} + \alpha) - \log\Gamma(\alpha)\right] + \log\Gamma(R\alpha) - \log\Gamma(O_d + R\alpha),$$

whence

$$\frac{\partial}{\partial \alpha} \log p(z_d \mid \alpha) = \sum_r \left[\Psi(O_{dr} + \alpha) - \Psi(\alpha)\right] + R \cdot \left[\Psi(R\alpha) - \Psi(O_d + R\alpha)\right] =: \hat{g}_\alpha(z_d),$$

where O_{dr} denotes the number of sentences in d assigned to relation r and $O_d = \sum_r O_{dr}$. We thus obtain a stochastic gradient

$$\hat{g}_\alpha = \frac{D}{S} \cdot \left[\sum_{d \in \mathcal{M}} \frac{1}{S'} \sum_{s'=1}^{S'} \hat{g}_\alpha\left(z_d^{(s')}\right)\right].$$

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

Matthew D. Hoffman, David M. Blei, and Francis R. Bach. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 856–864.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534.

David M. Mimno, Matthew D. Hoffman, and David M. Blei. 2012. Sparse stochastic inference for latent dirichlet allocation. In *ICML*.

John William Paisley, David M. Blei, and Michael I. Jordan. 2012. Variational bayesian inference with stochastic search. In *ICML*.

Rajesh Ranganath, Chong Wang, David M. Blei, and Eric P. Xing. 2013. An adaptive learning rate for stochastic variational inference. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 298–306.

Ellen Vorhees and David Graff. 2008. AQUAINT-2 Information-Retrieval Text Research Collection LDC2008T25. Web download. Philadelphia: Linguistic Data Consortium.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *EMNLP*, pages 1456–1466.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *ACL (1)*, pages 712–720.