Minibatch variational inference for various random graph models

Nils Bertschinger

November 12, 2019

Some thoughts on variational inference for the stochastic block model and other random graph models ...

1 Variational inference

Variational Bayes approximates the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ using the following identities

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \int q(\boldsymbol{\theta}) \ln p(\boldsymbol{x}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x},\boldsymbol{\theta}) q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x}) q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})} d\boldsymbol{\theta} \\
&= \mathbb{E}_q[\log p(\boldsymbol{x},\boldsymbol{\theta})] + H(q) + D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})) \\
&\geq \mathbb{E}_q[\log p(\boldsymbol{x},\boldsymbol{\theta})] + H(q) \; .
\end{aligned}$$

This last term is famously known as the *evidence lower bound* (ELBO) and commonly optimized wrt $q(\boldsymbol{\theta})$ in variational Bayes. It can also be written as

$$\begin{aligned}
&\mathbb{E}_q[\log p(\boldsymbol{x},\boldsymbol{\theta})] + H(q) \\
&\qquad = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})] + \mathbb{E}_q[-\log q(\boldsymbol{\theta})] \\
&\qquad = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{\theta}) - D_{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))
\end{aligned}$$

which is sometimes more convenient.

1.1 Stochastic gradient

Using the so called *reparametrization trick* the ELBO is amenable to stochastic gradient ascent. Consider a distribution $q(\boldsymbol{\theta};\boldsymbol{\eta})$ with variational parameters $\boldsymbol{\eta}$. Further, assume that a sample from q can be obtained by transforming

standard uniform or standard normal random variates $\boldsymbol{z}$, i.e. via some function $\boldsymbol{\theta} = f_{\boldsymbol{\theta}}(\boldsymbol{z}; \boldsymbol{\eta})$. Then, by Leibniz rule and the chain rule the derivative of the first term of the ELBO is given by

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\eta}} \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{\theta})] &= \frac{\partial}{\partial \boldsymbol{\eta}} \int q(\boldsymbol{\theta}; \boldsymbol{\eta}) \log p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{\partial}{\partial \boldsymbol{\eta}} \int p(\boldsymbol{z}) \log p(\boldsymbol{x}, f_{\boldsymbol{\theta}}(\boldsymbol{z}; \boldsymbol{\eta})) d\boldsymbol{z} \\ &= \mathbb{E}_{\boldsymbol{z}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\eta}} f_{\boldsymbol{\theta}}(\boldsymbol{z}; \boldsymbol{\eta}) \right]\end{aligned}$$

which is then stochastically evaluated using a few Monte-Carlo samples, i.e.

$$\frac{\partial}{\partial \boldsymbol{\eta}} \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{\theta})] \approx \frac{1}{M} \sum_i \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\eta}} f_{\boldsymbol{\theta}}(\boldsymbol{z}_i; \boldsymbol{\eta}) \ .$$

Note that also the gradient of the entropy $H(q)$ can be estimated via Monte-Carlo samples as

$$\begin{aligned}H(q(\boldsymbol{\theta})) &= \mathbb{E}_{q(\boldsymbol{\theta})}[-\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{\boldsymbol{z}}[-\log q(f_{\boldsymbol{\theta}}(\boldsymbol{z}))]\end{aligned}$$

where – in order to reduce the variance of the estimate – the same random samples of $\boldsymbol{z}$ should be used as for the other part of the ELBO. Indeed, we have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}}[\log p(\boldsymbol{x}, f_{\boldsymbol{\theta}}(\boldsymbol{z})) - \log q(f_{\boldsymbol{\theta}}(\boldsymbol{z}))] \\ &\qquad \approx \frac{1}{M} \sum_i \left(\log p(\boldsymbol{x}) + \log p(f_{\boldsymbol{\theta}}(\boldsymbol{z}_i) | \boldsymbol{x}) - \log q(f_{\boldsymbol{\theta}}(\boldsymbol{z}_i)) \right)\end{aligned}$$

has zero variance if the posterior approximation is exact, i.e. $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})$, and the same random samples $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ are used!

2 Random graph models

2.1 Stochastic block model

The *stochastic block model* (SBM) is a random graph model able to describe clustering/modularity and other structures in dense graphs. It is a special case of the more general class of graphin models that can be derived from exchangeability arguments [**?**].

According to the SBM a graph on N nodes, actually its adjacency matrix, is produced by the following generative process:

- Randomly assign a type $c_i \in \{1, \ldots, K\}$ to each node i:

$$c_i \sim \texttt{Categorical}(\theta_1, \ldots, \theta_K)$$

  Note: This assignment is unobserved, i.e. the SBM is an example of a *latent/hidden variable model*.

- Draw a link between i and j with probability $p_{c_i c_j}$, i.e.

$$A_{ij} \sim \texttt{Bernoulli}(p_{c_i c_j})$$

  Note: Links between nodes of the same types are independent (as in the Erdős-Rényi model).

Link probabilities are usually collected in a matrix $\boldsymbol{B} \in [0,1]^{K \times K}$, i.e. $(B)_{cc'} = p_{cc'}$ where $K \ll N$.

Thus, the parameters of the SBM are $\boldsymbol{\theta}$ and $\boldsymbol{B}$. In Bayesian statistics, we need to specify their prior distribution $p(\boldsymbol{\theta}, \boldsymbol{B})$. Commonly an independent Dirichlet prior for $\boldsymbol{\theta}$ and Beta priors for the entries of $\boldsymbol{B}$ are assumed. These are also the (conditionally) conjugate priors for the SBM.

Overall, we arrive at the following joint probability:

$$\begin{aligned}
p(\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B}) &= p(\boldsymbol{\theta}) p(\boldsymbol{B}) p(\boldsymbol{c}|\boldsymbol{\theta}) p(\boldsymbol{A}|\boldsymbol{c}, \boldsymbol{B}) \\
&= p(\boldsymbol{\theta}) \prod_{c,c'} p(B_{cc'}) \prod_i p(c_i|\boldsymbol{\theta}) \prod_{ij} p(A_{ij}|B_{c_i c_j}) \\
&= p(\boldsymbol{\theta}) \prod_{c,c'} p(B_{cc'}) \prod_i \theta_{c_i} \prod_{ij} B_{c_i c_j}^{A_{ij}} (1 - B_{c_i c_j})^{1 - A_{ij}} \\
&= p(\boldsymbol{\theta}) \prod_{c,c'} p(B_{cc'}) \prod_c \theta_c^{N_c} \prod_{c,c'} B_{cc'}^{n_{cc'}} (1 - B_{cc'})^{N_c N_{c'} - n_{cc'}}
\end{aligned}$$

where we have used the sufficient statistics

- N_c, counting number of nodes in each group c
- and $n_{cc'}$, counting number of edges between groups c and c'.

Note that it is not possible to marginalize over the latent class assignment, as the resulting summation

$$\sum_{\boldsymbol{c}} p(\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B}) = \sum_{c_1=1}^{K} \cdots \sum_{c_N=1}^{K} p(\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B})$$

involves exponentially many terms and cannot be simplified/separated.

Thus, we need to resort to some approximation algorithm ...

2.1.1 Variational inference

A tractable approximation for the SBM can be obtained by variational inference. In this case, the full posterior $p(\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B}|\boldsymbol{A})$ is approximated with simpler distribution that in particular factorizes over all nodes, i.e.

$$q(\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B}) = \prod_i q(c_i) q(\boldsymbol{\theta}) \prod_{c,c'} q(B_{cc'}) \; .$$

Here, the discrete distributions $q(c_i)$ are obviously categorical, i.e. $q(c; \boldsymbol{\eta}) = \eta_c$ and by conjugacy the optimal choice for $q(\boldsymbol{\theta})$ and $q(B_{cc'})$ can be shown to be the Dirichlet and Beta distribution respectively.

Using that approximation the log marginal likelihood can be approximated by the ELBO and especially the expectation over the approximating distribution for the latent class assignment can be carried out explicitly:

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{c},\boldsymbol{\theta},\boldsymbol{B})}[\log p(\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B})] + H(q(\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B})) \\
&= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{B})} \left[\mathbb{E}_{q(\boldsymbol{c})} \left[\log p(\boldsymbol{\theta}) + \sum_{cc'} \log p(B_{cc'}) + \sum_i \log p(c_i|\boldsymbol{\theta}) + \sum_{ij} \log p(A_{ij}|B_{c_i c_j}) \;\middle|\; \boldsymbol{\theta}, \boldsymbol{B} \right] \right] + H(q(\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{B})) \\
&= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{B})} \left[\log p(\boldsymbol{\theta}) + \sum_{cc'} \log p(B_{cc'}) + \sum_{c_1=1}^{K} \cdots \sum_{c_N=1}^{K} \prod_i q(c_i) \left(\sum_i \log p(c_i|\boldsymbol{\theta}) + \sum_{ij} \log p(A_{ij}|B_{c_i c_j}) \right) \right] + H \\
&= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{B})} \left[\log p(\boldsymbol{\theta}) + \sum_{cc'} \log p(B_{cc'}) + \sum_i \sum_{c_i=1}^{K} q(c_i) \log p(c_i|\boldsymbol{\theta}) + \sum_{ij} \sum_{c_i=1}^{K} \sum_{c_j=1}^{K} q(c_i) q(c_j) \log p(A_{ij}|B_{c_i c_j}) \right] \\
&\quad + \sum_i H(q(c_i)) + H(\boldsymbol{\theta}) + \sum_{cc'} H(q(B_{cc'})) \\
&= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{B})} \left[\log p(\boldsymbol{\theta}) + \sum_{cc'} \log p(B_{cc'}) + \sum_i \sum_{c_i=1}^{K} \eta_{c_i} \log \theta_{c_i} + \sum_{ij} \sum_{c_i=1}^{K} \sum_{c_j=1}^{K} \eta_{c_i} \eta_{c_j} (A_{ij} \log B_{c_i c_j} + (1 - A_{ij}) \log \right. \\
&\quad \left. - \sum_i \sum_{c_i=1}^{K} \eta_{c_i} \log \eta_{c_i} + \mathbb{E}_{q(\boldsymbol{\theta})}[-\log q(\boldsymbol{\theta})] + \sum_{cc'} \mathbb{E}_{q(B_{cc'})}[-\log q(B_{cc'})] \right.
\end{aligned}$$

The remaining expectation over $q(\boldsymbol{\theta}, \boldsymbol{B})$ can be approximated by Monte-Carlo sampling as explained above. Further, the sum $\sum_{c_i=1}^{K} \sum_{c_j=1}^{K} \eta_{c_i} \eta_{c_j} \log B_{c_i c_j}$ can be conveniently written as a matrix multiplication $\boldsymbol{\eta}_i^T \log(\boldsymbol{B}) \boldsymbol{\eta}_j$.

2.1.2 Scalable algorithm

A scalable algorithm is then obtained by maximizing the stochastic ELBO with respect to all variational parameters, i.e. of $q(\boldsymbol{\theta}, \boldsymbol{B})$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N$.

For even larger data sets also minibatch training should be feasible. To this

end, the ELBO can be written as a sum over all vertex pairs

$$\sum_{\substack{ij \\ i \neq j}} \Bigg[\frac{1}{N(N-1)} \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] + \frac{1}{N(N-1)} \sum_{cc'} \mathbb{E}_{q(B_{cc'})}[\log p(B_{cc'}) - \log q(B_{cc'})] + \frac{1}{N-1} \sum_{c_i=1}^{K} \eta_{c_i} (\mathbb{E}_{q(\boldsymbol{\theta})}[\log \theta_{c_i}] - \log \eta_{c_i}) + \sum_{c_i=1}^{K} \sum_{c_j=1}^{K} \eta_{c_i} \eta_{c_j} \mathbb{E}_{q(B_{c_i c_j})}[A_{ij} \log B_{c_i c_j} + (1 - A_{ij}) \log(1 - B_{c_i c_j})] \Bigg]$$

where we assume that the diagonal of the adjacency matrix, i.e. $i = j$, is excluded from the likelihood. The gradient can then be evaluated with respect to a minibatch containing just a fraction of all $N(N-1)$ vertex pairs.

Note that using conjugate priors and matches variational distributions all expectations can be computed analytically. The minibatch ELBO is thus a function of the variational parameters and can be optimized with standard optimizers, e.g. minibatch stochastic gradient ascent or BFGS over all data points. Obviously, using the reparameterization explained above also a doubly stochastic variant with additional Monte-Carlo sampling of the expectations is feasible.

2.2 Deep graphon model

The SBM is actually a special case of more general graphon models. These can be derived from the Aldous-Hover theorem, stating that an infinitely exchangeable random array A can be generated as

$$\begin{aligned} u_i &\sim \texttt{Uniform}(0,1) \quad \forall i = 1, \ldots, N \\ A_{ij} &\sim \texttt{Bernoulli}(W(u_i, u_j)) \end{aligned}$$

with the *graphon* $W : [0,1]^2 \to [0,1]$ as the main parameter capturing the structure of the random graph. Note that the SBM can be represented by a piecewise constant graphon

$$W_{\text{SBM}}(u, v) = B_{c_u, c_v}$$

where $c_u = \text{argmin} \left\{ i \in \{1, \ldots, K\} | u < \sum_k^i \theta_k \right\}$. Then, for $u \sim \texttt{Uniform}(0,1)$ we have that $c_u \sim \texttt{Categorical}(\theta_1, \ldots, \theta_K)$.

Many other choices for the graphon W are possible and have been proposed in the literature [**?**]. In general, the graphon is only identified up to measure preserving transformations – akin to the combinatorial symmetry of label switching in the SBM. The model can be identified by either imposing a strong prior, e.g. enforcing smoothness of W, or using a canonical representation of the graphon,

e.g. imposing that the degree density $g(u) = \int_0^1 W(u,v)dv$ is (strictly) increasing in u.

Completing the model with a suitable prior on W, the joint probability can be written as

$$\begin{aligned} p(\boldsymbol{A}, \boldsymbol{u}, W) &= p(W)p(\boldsymbol{u})p(\boldsymbol{A}|\boldsymbol{u}, W) \\ &= p(W)\prod_i p(u_i)\prod_{ij} p(A_{ij}|W(u_i,u_j)) \\ &= p(W)\prod_{ij} W(u_i,u_j)^{A_{ij}}(1-W(u_i,u_j))^{1-A_{ij}} \end{aligned}$$

where we have used that $p(u_i) = 1$.

As before, the marginal probability $p(\boldsymbol{A}, W)$ is an intractable integral over the N-dimensional unit cube

$$p(\boldsymbol{A}, W) = \int_0^1 \cdots \int_0^1 p(\boldsymbol{A}, \boldsymbol{u}, W)du_1 \dots du_N$$

and we resort to a variational approximation again.

2.2.1 Variational inference

As before, the full posterior $p(\boldsymbol{u}, W|\boldsymbol{A})$ can be approximated with a product distribution

$$q(\boldsymbol{u}, W) = \prod_i q(u_i)q(W) \ .$$

Without further assumptions on W the optimal distribution for $q(u_i)$ is not known and any convenient choice for it can be assumed, e.g. a logit-normal distribution. Further, the resulting expectation in the ELBO over the continuous latent variables $\boldsymbol{u}$ cannot be carried out analytically. Instead, the ELBO is amenable to stochastic gradient ascent using Monte-Carlo sampling as explained in section 1.1. Overall, we obtain

$$\begin{aligned} &\mathbb{E}_{q(\boldsymbol{u})q(W)}[\log p(\boldsymbol{A}, \boldsymbol{u}, W)] + H(q(\boldsymbol{u}, W)) \\ &= \mathbb{E}_{q(\boldsymbol{u})q(W)}[\log p(W) + \sum_i \log p(u_i) + \sum_{ij} p(A_{ij}|W(u_i,u_j))] + \sum_i H(q(\boldsymbol{u})) + H(q(W)) \\ &= \mathbb{E}_{q(\boldsymbol{u})q(W)}[\log p(W) - \log q(W) + \sum_i \log p(u_i) - \sum_i \log q(u_i) + \sum_{ij} p(A_{ij}|W(u_i,u_j))] \\ &= \mathbb{E}_{q(\boldsymbol{u})q(W)}[\log p(W) - \log q(W) - \sum_i \log q(u_i) + \sum_{ij} A_{ij}\log W(u_i,u_j) + (1-A_{ij})\log(1-W(u_i,u_j))] \ . \end{aligned}$$

2.2.2 Scalable algorithm

A scalable algorithm is then obtained by writing the ELBO as a sum over all vertex pairs

$$\sum_{\substack{ij \\ i\neq j}} \mathbb{E}_{q(\boldsymbol{u})q(W)} \Bigg[\frac{1}{N(N-1)} (\log p(W) - \log q(W)) \\ - \frac{1}{N-1} \log q(u_i) + A_{ij} \log W(u_i, u_j) + (1 - A_{ij}) \log(1 - W(u_i, u_j)) \Bigg]$$

and approximating the expectations over the variational distributions with few Monte-Carlo samples, i.e.

$$\sum_{\substack{ij \\ i\neq j}} \frac{1}{M} \sum_k \Bigg[\frac{1}{N(N-1)} (\log p(W^k) - \log q(W^k)) \\ - \frac{1}{N-1} \log q(u_i^k) + A_{ij} \log W(u_i^k, u_j^k) + (1 - A_{ij}) \log(1 - W(u_i^k, u_j^k)) \Bigg]$$

where $W^k \sim q(W), u_i^k \sim q(u_i)$. Using the reparameterization trick a doubly stochastic gradient ascent method can be obtained.

Note that in contrast to the SBM above, we have not assumed that the entropy of the variational distributions is available in closed form. This is indeed not the case for the logit-normal distribution which might be a natural choice for $q(u; \mu_u, \sigma_u)$ as it is easily reparameterized as

$$z \sim \texttt{Normal}(0, 1)$$
$$u = \sigma(\mu_u + \sigma_u z)$$

with variational parameters μ_u, σ_u and the logistic sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$. By change of measure its log density is

$$\begin{aligned} \log q(u) &= \log \texttt{Normal}(\sigma^{-1}(u) | \mu_u, \sigma_u) + \log |\frac{d\sigma^{-1}(u)}{du}|] \\ &= \log \texttt{Normal}(\sigma^{-1}(u) | \mu_u, \sigma_u) - \log u - log(1-u) \end{aligned}$$

as $\frac{d\sigma^{-1}(u)}{du} = \frac{1}{\frac{du}{dx}}$ and $\frac{du}{dx} = \frac{d\sigma(x)}{dx} = (\sigma(x)(1-\sigma(x)) = u(1-u)$. Correspondingly, its entropy can be partially derived as

$$H(q(u)) = H(\texttt{Normal}(\mu_u, \sigma_u)) - \mathbb{E}_x[\log \sigma(x) + \log(1 - \sigma(x))]$$

where $x \sim \texttt{Normal}(\mu_u, \sigma_u)$.

2.3 Geometric Inhomogeneous Random Graphs

Geometric random graphs are also latent variables models where each vertex $v \in V$ is assigned a position x_v in some metric space, e.g. the hyperbolic

plane. The probability of an edge between vertices u, v is then derived via some function of the metric distance between x_u and x_v. Recently, the *geometric inhomogeneous random graph* (GIRG) model has been proposed and received some interest as it includes the *hyperbolic random graph* model as a special case.

The GIRG model assigns two latent variables to each vertex v, a weight w_v as well as position $\boldsymbol{x}_v$. Its generative story reads as follows

$$\begin{aligned} w_i &\sim \texttt{PowerLaw}(\beta) \\ \boldsymbol{x}_i &\sim \texttt{Uniform}(\mathbb{T}^d) \\ A_{ij} &\sim \texttt{Bernoulli}\left(c \cdot \min\left\{\frac{1}{||\boldsymbol{x}_i - \boldsymbol{x}_j||^{\alpha d}}\left(\frac{w_i w_j}{W}\right)^{\alpha}, 1\right\}\right) \end{aligned}$$

where $\mathbb{T}^d$ denotes the d-dimensional torus and distances are taken as the ∞-norm on $\mathbb{T}^d$, i.e. $||\boldsymbol{x} - \boldsymbol{y}|| = \max_d |x_d - y_d|_C$ with $|x - y|_C = \min\{|x - y|, 1 - |x - y|\}$. Further, $W = \sum_i w_i$ and $\alpha > 1, \beta > 2$ and $c > 0$ are parameters of the model.

According to [**?**] the hyperbolic random graph model can be obtained by $d = 1 \ldots$

- The model seems to approximate a power law distribution ...
- Also the mapping of the distance function is not exact?

Given that it might make sense to consider the hyperbolic model in its own right ...

2.3.1 Variational inference

Again, we complete the model with suitable priors for all parameters and assume a factor distribution approximation to the full posterior, i.e.

$$q(\alpha, \beta, c, \boldsymbol{w}, \boldsymbol{x}) = q(\alpha)q(\beta)q(c)\prod_i q(w_i)\prod_i q(\boldsymbol{x}_i)\ .$$

The ELBO then reads

$$\begin{aligned} &\mathbb{E}_{q(\alpha,\beta,c,\boldsymbol{w},\boldsymbol{x})}[\log p(\boldsymbol{A}, \alpha, \beta, c, \boldsymbol{w}, \boldsymbol{x})] \\ &= \mathbb{E}_{q(\alpha,\beta,c,\boldsymbol{w},\boldsymbol{x})}[\log p(\alpha) + \log p(\beta) + \log p(c) + \sum_i \log p(w_i) + \sum_i \log p(\boldsymbol{x}_i) + \sum_{ij} \log \texttt{Bernoulli}(A_{ij}|p_{ij})] \end{aligned}$$

where $\log p(\boldsymbol{x}_i) \equiv 0$ and $p_{ij} = c \cdot \min\left\{\frac{1}{||\boldsymbol{x}_i - \boldsymbol{x}_j||^{\alpha d}}\left(\frac{w_i w_j}{W}\right)^{\alpha}, 1\right\}$.

2.3.2 Scalable algorithm

As before, pulling out the summation over all vertex pairs we obtain the mini-batch ELBO

$$\sum_{\substack{ij \\ i \neq j}} \mathbb{E}_{q(\alpha,\beta,c,\boldsymbol{w},\boldsymbol{x})} \Bigg[\frac{1}{N(N-1)} \log p(\alpha) + \frac{1}{N(N-1)} \log p(\beta) + \frac{1}{N(N-1)} \log p(c) + \frac{1}{(N-1)} \log p(w_i) + \log \texttt{Bernoulli}(A_{ij} | p_{ij}) \Bigg]$$

which should be optimizable via doubly stochastic gradient ascent.

3 And beyond ...