Now, the derivative of $\mathcal{L}_{obj}$ with respect to each entry $\hat{\rho}_m$ of the vector $\hat{\rho}$, for $m \in 1, 2, \dots K$, is:

$$\begin{aligned}\frac{\partial \mathcal{L}_{obj}}{\partial \hat{\rho}_m} &= \hat{\omega}_m \big(K + 1 - \hat{\rho}_m \hat{\omega}_m\big) \psi_1(\hat{\rho}_m \hat{\omega}_m) \\ &\quad - \hat{\omega}_m \big(K(K+1-m) + 1 + \gamma - (1 - \hat{\rho}_m)\hat{\omega}_m\big) \psi_1((1-\hat{\rho}_m)\hat{\omega}_m) \\ &\quad + \textstyle\sum_{\ell=1}^{K} \Delta_{m\ell} \bigg(\log(\alpha + \kappa) - \log \kappa + \textstyle\sum_{k=0}^{K} \alpha_k P_{k\ell}(\hat{\theta}) \bigg) \\ &\quad + \Delta_{m,K+1} \big(\sum_{k=0}^{K} \alpha_k P_{k,K+1}(\hat{\theta}) \big)\end{aligned} \tag{48}$$

B.2 Unconstrained Optimization Problem

In practice, we find that it is more numerically stable to first transform our constrained optimization problem above into an unconstrained problem, and then solve the unconstrained problem via a modern gradient descent algorithm (L-BFGS).

Our transformation follows exactly the steps outlined in the Supplement of [11], where the constrained objective for the HDP topic model was transformed into unconstrained variables. See that Supplement document for mathematical details, or our public source code[7] for practical details.

C Efficient updates for the HDP-aMMSB

This section provides details of how to compute updates for the HDP-aMMSB in $O(N^2K)$ time, as opposed to the naive $O(N^2K^2)$ time. This method was first noted in [14]; we provide a single, cohesive explanation here. The key observation is that we only need to compute the N^2K diagonal entries of $\hat{\eta}$ using Eq. 27; that is, we only compute and store $\hat{\eta}_{ij\ell m}$ for $\ell = m$. We also need to compute and store Z_{ij}, the normalization constant that makes $\hat{\eta}_{ij}$ sum to 1. We can do so in $O(K)$ time by noting:

$$\begin{aligned}Z_{ij} \triangleq \sum_{\ell,m}^{K} \hat{\eta}_{ij\ell m} &= \sum_{k}^{K} \tilde{\pi}_{ij} \tilde{\pi}_{jk} f(\phi_k, x_{ij}) + \sum_{\ell \neq m}^{K} \tilde{\pi}_{i\ell} \tilde{\pi}_{jm} f(\varepsilon, x_{ij}) \\ &= \sum_{k}^{K} \tilde{\pi}_{ik} \tilde{\pi}_{jk} \Big(f(\phi_k, x_{ij}) - f(\varepsilon, x_{ij}) \Big) + \tilde{\pi}_i \tilde{\pi}_j f(\varepsilon, x_{ij})\end{aligned} \tag{49}$$

where we have substituted in Eq. 27 in for $\hat{\eta}_{ij\ell m}$. This restricted computation of $\hat{\eta}$ makes updates to $q(\pi)$ and computation of $\mathcal{L}$ slightly more complicated. We outline the relevant terms below. A helpful identity that follows from the definition of $\hat{\eta}$ and $\tilde{\pi}$ is:

$$\sum_{\substack{m \\ m \neq \ell}}^{K} \hat{\eta}_{ij\ell m} = \frac{1}{Z_{ij}} f(\varepsilon, x_{ij}) \tilde{\pi}_{i\ell} (\tilde{\pi}_j - \tilde{\pi}_{j\ell}) \tag{50}$$

Updates to $q(\pi)$. From Eq. 28, we have that the usage sufficient statistic for the HDP-aMMSB is $U_{i\ell} = \sum_{j=1}^{N} \sum_{m=1}^{K} (\hat{\eta}_{ij\ell m} + \hat{\eta}_{jim\ell})$ We can compute $U_{i\ell}$ given only the diagonal $\hat{\eta}_{ij\ell\ell}$ and normalization constants by:

$$U_{i\ell} = \sum_{j}^{N} \Big(\hat{\eta}_{ij\ell\ell} + \frac{1}{Z_{ij}} \tilde{\pi}_{i\ell} (\tilde{\pi}_j - \tilde{\pi}_{j\ell}) \Big) + \sum_{j}^{N} \Big(\hat{\eta}_{ji\ell\ell} + \frac{1}{Z_{ji}} \tilde{\pi}_{i\ell} (\tilde{\pi}_j - \tilde{\pi}_{j\ell}) \Big) \tag{51}$$

[7] https://bitbucket.org/michaelchughes/bnpy