Supplementary Materials for
Complexity of Inference in Latent Dirichlet Allocation

A Proof of Lemma 2

Proof of Lemma 2. Assume there are T sets each having $k \geq 3$ elements, and let Φ be the optimal LDA objective. Define $F(n) = \log \Gamma(n + \alpha)$. Since l_{it} is constant across all topics, the linear term in Eq. 2 will be a constant K. First, note that, if there is a perfect matching,

$$\Phi \geq \frac{n}{k} F(k) + (T - \frac{n}{k}) F(0) + K. \tag{16}$$

The $F(0)$ term is the contribution of unused topics. Otherwise, assume that the best packing has $\gamma \leq cn/k$ sets, each with k elements. Then, by the properties of the log-gamma function,

$$\Phi \leq \gamma F(k) + \frac{n - \gamma k}{k-1} F(k-1) + (T - \frac{n}{k}) F(0) + K, \tag{17}$$

where we assume, conservatively, that all of the remaining words are explained by topics assigned $(k-1)$ words. Also, since there was no perfect matching, there were at most $T - \frac{n}{k}$ unused topics. Using our bound on γ, we have

$$\begin{aligned}
\Phi &\leq \frac{cn}{k} F(k) + \frac{n - \frac{cn}{k} k}{k-1} F(k-1) && + (T - \frac{n}{k}) F(0) + K && (18)\\
&= \frac{cn}{k} F(k) + \frac{n(1-c)}{k-1} F(k-1) && + (T - \frac{n}{k}) F(0) + K && (19)\\
&= \frac{dn}{k} F(k) && + (T - \frac{n}{k}) F(0) + K, && (20)
\end{aligned}$$

where

$$d := c + (1-c)\beta, \qquad \text{for } \beta := \frac{k}{F(k)} \frac{F(k-1)}{k-1}. \tag{21}$$

Note that $F(k)/k \to \infty$ as $k \to \infty$. Along with the convexity of F, it follows that there exists a k_0 such that $\beta < 1$ for all $k > k_0$. Note that $k > (3+\alpha)^2$ suffices. This implies that $d < 1$, which shows that there is a non-zero gap between the possible values of Φ. $\square$

Note that the maximum concentration objective, $F(n) = n \log n$, satisfies the conditions on F and, in particular, we have $\beta < 1$ for $k = 3$.

B Derivation of MAP θ objective

$$\begin{aligned}
\Pr(\theta|\mathbf{w}) &\propto \sum_{z_1,\ldots,z_N} \Pr(\theta) \Pr(w_1, \ldots, w_N, z_1, \ldots, z_N|\theta) && (22)\\
&= \Pr(\theta) \sum_{z_1,\ldots,z_N} \prod_{i=1}^{N} \Pr(z_i|\theta) \Pr(w_i|z_i) && (23)\\
&= \Pr(\theta) \prod_{i=1}^{N} \sum_{z_i} \Pr(z_i|\theta) \Pr(w_i|z_i) && (24)\\
&= \Pr(\theta) \prod_{i=1}^{N} \sum_{t=1}^{T} \theta_t \Pr(w_i|z_i = t) && (25)\\
&\propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1} \prod_{i=1}^{N} \sum_{t=1}^{T} \theta_t \Pr(w_i|z_i = t). && (26)
\end{aligned}$$

C Proof of Lemma 3

If $\epsilon \geq K(\alpha, T, N)$ the claim trivially holds. Assume for the purpose of contradiction that there exists a word $\hat{i}$ such that $\theta^*_{\hat{t}} < K(\alpha, T, N)$, where $\hat{t} = \arg\max_t \psi_{\hat{i}t}\theta^*_t$.

Let Y denote the set of topics $t \neq \hat{t}$ such that $\theta^*_t \geq 2\epsilon$. Let $\beta_1 = \sum_{t\in Y} \theta^*_t$ and $\beta_2 = \sum_{t\notin Y, t\neq \hat{t}} \theta^*_t$. Note that $\beta_2 < 2T\epsilon$. Consider $\hat{\theta}$ defined as follows:

$$\hat{\theta}_{\hat{t}} = \frac{1}{N} \tag{27}$$

$$\hat{\theta}_t = \left(\frac{1-\beta_2-\frac{1}{N}}{\beta_1}\right)\theta^*_t \text{ for } t \in Y \tag{28}$$

$$\hat{\theta}_t = \theta^*_t \text{ for } t \notin Y, t \neq \hat{t}. \tag{29}$$

Note that this construction implies the bound $\hat{\theta}_t \geq \left(1 - 2T\epsilon - \frac{1}{N}\right)\theta^*_t$ for $t \in Y$. Assuming $n \geq 4$ and $\epsilon \leq \frac{1}{2TN}$, we have that $\hat{\theta}_t \geq \frac{1}{2}\theta^*_t \geq \epsilon$ for $t \in Y$, so $\hat{\theta}$ is feasible.

We will show that $\Phi(\hat{\theta}) > \Phi(\theta^*)$, contradicting the optimality of θ^*. First we need the following upper bound, which uses the fact that $\theta^*_{\hat{t}} < \frac{1}{N}$:

$$\frac{1-\beta_2-\frac{1}{n}}{\beta_1} = \frac{1-\beta_2-\frac{1}{N}}{1-\beta_2-\theta^*_{\hat{t}}} \tag{30}$$

$$< 1. \tag{31}$$

Then, we have:

$$\frac{P(\hat{\theta})}{\alpha-1} = \sum_{t=1}^{T}\log(\hat{\theta}_t) \tag{32}$$

$$= \log\frac{1}{N} + \sum_{t\in Y}\log\left(\frac{1-\beta_2-\frac{1}{N}}{\beta_1}\right) + \sum_{t\neq\hat{t}}\log\theta^*_t \tag{33}$$

$$\leq \log\frac{1}{N} + \sum_{t\neq\hat{t}}\log\theta^*_t. \tag{34}$$

Thus,

$$\frac{P(\hat{\theta})-P(\theta^*)}{\alpha-1} \leq \log\frac{1}{N} - \log\theta^*_{\hat{t}} \tag{35}$$

which when $\alpha < 1$ gives the inequality:

$$P(\hat{\theta})-P(\theta^*) \geq (\alpha-1)\Big(\log\frac{1}{N} - \log\theta^*_{\hat{t}}\Big) \tag{36}$$

$$= (1-\alpha)\Big(\log N + \log\theta^*_{\hat{t}}\Big). \tag{37}$$

Moving on to the second term, we have:

$$L(\hat{\theta}) = \sum_{j\in[N]\,:\,j\neq\hat{i}}\log\left(\sum_t \psi_{jt}\hat{\theta}_t\right) + \log\left(\sum_t \psi_{\hat{i}t}\hat{\theta}_t\right) \tag{38}$$

$$\geq (N-1)\log\left(\frac{1-\beta_2-\frac{1}{N}}{\beta_1}\right) + \sum_{j\in[N]\,:\,j\neq\hat{i}}\log\left(\sum_t \psi_{jt}\theta^*_t\right) + \log\left(\frac{\psi_{\hat{i}\hat{t}}}{N}\right) \tag{39}$$

$$\geq (N-1)\log\left(1-\frac{2}{N}\right) + \sum_{j\in[N]\,:\,j\neq\hat{i}}\log\left(\sum_t \psi_{jt}\theta^*_t\right) + \log\left(\frac{\psi_{\hat{i}\hat{t}}}{N}\right). \tag{40}$$

$$
\begin{aligned}
L(\hat{\theta}) - L(\theta^*) &\geq (N-1)\log\left(1 - \frac{2}{N}\right) + \log\left(\frac{\psi_{\hat{i}\hat{t}}}{N}\right) - \log\left(\sum_t \psi_{\hat{i}t}\theta^*_t\right) && (41)\\
&\geq (N-1)\log\left(1 - \frac{2}{N}\right) + \log\left(\frac{\psi_{\hat{i}\hat{t}}}{N}\right) - \log\left(T\psi_{\hat{i}\hat{t}}\theta^*_{\hat{t}}\right) && (42)\\
&= (N-1)\left(\log(N-2) - \log N\right) + \log\left(\frac{1}{N}\right) - \log\left(T\theta^*_{\hat{t}}\right) && (43)\\
&\geq (N-1)\left(-\frac{2}{N-2}\right) + \log\left(\frac{1}{N}\right) - \log\left(T\theta^*_{\hat{t}}\right) && (44)\\
&\geq -3 + \log\left(\frac{1}{N}\right) - \log\left(T\theta^*_{\hat{t}}\right). && (45)
\end{aligned}
$$

where we used the lower bound $\log(N-2) \geq \log N - \frac{2}{N-2}$ that arises from the convexity of the $\log(x)$ function, and again assumed $N \geq 4$.

Finally, putting these two together, we have:

$$\Phi(\hat{\theta}) - \Phi(\theta^*) \geq -3 - \alpha \log N - \log T - \alpha \log \theta^*_{\hat{t}} \tag{46}$$

Plugging in $\theta^*_{\hat{t}} < K(\alpha, T, N)$ results in $\Phi(\hat{\theta}) - \Phi(\theta^*) > 0$, giving the contradiction.

D Proof of Theorem 8

Here we reduce from the *unique* set cover problem, where we are guaranteed that there is only one minimal size set that covers all elements. It can be shown that Unique Set Cover is NP-hard (under randomized reductions) by using standard reductions from Unique SAT to Vertex Cover, and then from Vertex Cover to Set Cover.

Consider a Unique Set Cover instance and our standard reduction to an LDA instance as described in earlier sections. In particular, let $\mathbf{w} = (w_1, \ldots, w_N)$ denote a Unique Set Cover reduction instance and let $C \subseteq [T]$ denote the unique minimum cover. Let S_i be those sets (topics) that cover element w_i. We will show that for sufficiently small hyperparameters α_t, we can determine whether a set (topic) $t \in C$ by testing the value of $\mathbb{E}[\theta_t|X]$, thus proving that the computation of the latter is NP-hard.

We have

$$
\begin{aligned}
p(\theta \mid X) &\propto \prod_t \theta_t^{\alpha_t - 1} \prod_i \sum_{t' \in S_i} \theta_{t'} && (47)\\
&= \sum_r \prod_t \theta_t^{\alpha_t - 1 + \eta_t(r)}, && (48)
\end{aligned}
$$

where the final summation is over elements $r \in \mathcal{R} := S_1 \times \cdots \times S_N$ and $\eta_t(r) := |\{i : r_i = t\}|$. For $r \in \mathcal{R}$, we write $|r|$ to denote the number of topics t such that $\eta_t(r) \neq 0$.

Let $\mathcal{N}$ denote the set of sequences $n = (n_1, \ldots, n_T)$ such that $n_t \geq 0$ and $\sum_t n_t = N$. For $n \in \mathcal{N}$, define

$$Z(n) := \int \cdots \int \prod_t \theta_t^{\alpha_t - 1 + n_t}\, d\theta_1 \cdots d\theta_T = \frac{\prod_t \Gamma(\alpha_t + n_t)}{\Gamma(\bar{\alpha} + N)}, \tag{49}$$

where $\bar{\alpha} = \sum_t \alpha_t$. It follows that

$$
\begin{aligned}
\mathbb{E}[\theta_t|X] &= \int \cdots \int \theta_t \cdot p(\theta \mid X) d\theta_1 \cdots d\theta_T && (50)\\
&= \frac{1}{\sum_r Z(\eta(r))} \int \cdots \int \theta_t \sum_r \prod_\tau \theta_\tau^{\alpha_\tau - 1 + n_\tau(r)} d\theta_1 \cdots d\theta_T && (51)\\
&= \frac{1}{\sum_r Z(\eta(r))} \sum_r Z(\eta(r, t)), && (52)
\end{aligned}
$$

where $\eta(r,t) \in \mathcal{N}$ is given by $\eta(r,t)_\tau = \eta_\tau(r) + 1(\tau = t)$. By the identity $\Gamma(z+1) = z\Gamma(z)$, we have $Z(\eta(r,t)) = Z(\eta(r))\frac{\alpha_t + n_t(r)}{\bar{\alpha}+N}$, and so it follows that

$$\mathbb{E}[\theta_t|X] = \sum_{r\in\mathcal{R}} \bar{Z}(\eta(r))\frac{\alpha_t + n_t(r)}{\bar{\alpha}+N}, \tag{53}$$

where

$$\bar{Z}(n) := \frac{Z(n)}{\sum_{r'\in\mathcal{R}} Z(\eta(r'))}. \tag{54}$$

For $c \in [T]$, let $\mathcal{R}_c$ denote the set of those $r \in \mathcal{R}$ such that $|r| = c$. Recall that $C \subseteq [T]$ is the unique minimum set cover associated with the reduction X. Thus, for $t \in C$,

$$\mathbb{E}[\theta_t|X] = \sum_{r\in\mathcal{R}_{|C|}} \bar{Z}(\eta(r))\frac{\alpha_t + n_t(r)}{\bar{\alpha}+N} + \sum_{r\in\mathcal{R}\setminus\mathcal{R}_{|C|}} \bar{Z}(\eta(r))\frac{\alpha_t + n_t(r)}{\bar{\alpha}+N} \tag{55}$$

$$\geq \frac{\alpha_t + 1}{\bar{\alpha}+N}\sum_{r\in\mathcal{R}_{|C|}} \bar{Z}(\eta(r)), \tag{56}$$

where we have used the observation that $\eta_t(r) \geq 1$ for $r \in \mathcal{R}_{|C|}$. Let $\beta = \sum_{r\in\mathcal{R}_{|C|}} \bar{Z}(\eta(r))$. If $t \notin C$, then $n_t(r) = 0$ for $r \in \mathcal{R}_{|C|}$ and so we have $\mathbb{E}[\theta_t|X] \leq \beta\frac{\alpha_t}{\bar{\alpha}+N} + (1-\beta)$. It follows that if

$$\beta > \frac{1}{2}\Big(1 + \frac{\bar{\alpha}+N}{\bar{\alpha}+N+1}\Big) \tag{57}$$

then the minimum cover C contains a topic t if and only if $\mathbb{E}[\theta|X] \geq \frac{1}{4}\Big(1 + 3\frac{\bar{\alpha}+N}{\bar{\alpha}+N+1}\Big)\frac{\alpha_t+1}{\bar{\alpha}+N}$, and, moreover, we can determine the minimum cover from a bound on β and polynomial approximations to the marginal distributions of the components of θ. We will take $\alpha_t = \alpha$ henceforth, and show that for α small enough, the bound (57) indeed holds.

Let $n, n' \in \mathcal{N}$ be topic counts associated with the minimal cover and some non-minimal cover, respectively. That is, let $n = \eta(r)$ for some $r \in \mathcal{R}_{|C|}$ and let $n' = \eta(r')$ for some $r' \in \mathcal{R}_k$ and $k > |C|$. We will bound $Z(n')/Z(n)$ in order to bound β. We have

$$\prod_t \Gamma(\alpha + n_t) \geq \Gamma(\alpha)^{T-|C|}\,\Gamma(\alpha+1)^{|C|}, \tag{58}$$

whereas

$$\prod_t \Gamma(\alpha + n'_t) \leq \Gamma(\alpha)^{T-|C|-1}\,\Gamma(\alpha+1)^{|C|}\,\Gamma(\alpha+N-|C|). \tag{59}$$

Therefore,

$$\frac{Z(n')}{Z(n)} \leq \frac{\Gamma(\alpha)^{T-|C|-1}\,\Gamma(\alpha+1)^{|C|}\,\Gamma(\alpha+N-|C|)}{\Gamma(\alpha)^{T-|C|}\,\Gamma(\alpha+1)^{|C|}} \tag{60}$$

$$= \frac{\Gamma(\alpha+N-|C|)}{\Gamma(\alpha)}. \tag{61}$$

By the convexity of $\Gamma(1/c)$ in c, we have $\Gamma(\alpha) \geq \alpha^{-1} - \gamma$, where $\gamma \approx .577$ is the Euler constant. Therefore,

$$\frac{Z(n')}{Z(n)} \leq \frac{\Gamma(\alpha+N-1)}{\alpha^{-1}-\gamma} =: \kappa(\alpha). \tag{62}$$

Then by conservatively assuming that there is only one responsibility corresponding to the minimum cover, we have that

$$\beta \geq \frac{Z(n)}{Z(n) + T^N\kappa(\alpha)Z(n)}. \tag{63}$$

Therefore, the bound (57) is achieved when

$$\kappa(\alpha) \leq \frac{1}{T^N(2\bar{\alpha} + 2N + 1)}. \tag{64}$$

In particular, when $N, T \geq 2$ and

$$\alpha^{-1} > 2T^N \Gamma(N)(2N + 2) \tag{65}$$

the marginal expectations can be used to read off the unique minimal set cover.