Gibbs Sampler for LDA

David Mareček

November 8, 2018
\[
p(z_{nd} = k|\{w\}, \{z_{-nd}\}, \gamma, \alpha) = \\
= \frac{p(z_{nd} = k|\{z_{-nd}\}, \gamma, \alpha) p(\{w\}|z_{nd} = k, \{z_{-nd}\}, \gamma, \alpha)}{p(\{w\}|\{z_{-nd}\}, \gamma, \alpha)} \\
\propto p(z_{nd} = k|\{z_{-nd}\}, \alpha) p(\{w\}|z_{nd} = k, \{z_{-nd}\}, \gamma) \\
\propto p(z_{nd} = k|\{z_{-nd}\}, \alpha) p(w_{nd}|\{w_{-nd}\}, z_{nd} = k, \{z_{-nd}\}, \gamma) \\
\propto \int p(z_{nd} = k|\theta_d) p(\theta_d|z_{-nd}, \alpha) d\theta_d \int p(w_{nd}|\beta_k) p(\beta_k|\{w_{-nd}\}, \{z_{-nd}\}, \gamma) d\beta_k \\
= \frac{\alpha + c^k_d}{K\alpha + N_d - 1} \frac{\gamma + \tilde{c}^k_{w_{nd}}}{M\gamma + \sum_{m=1}^{M} \tilde{c}^k_m}
\]

- \(c^k_d\) . . . How many words in document \(d\) are assigned to topic \(k\).
- \(\tilde{c}^k_m\) . . . How many times the word \(m\) is assigned to topic \(k\) (across all documents).
**LDA Algorithm**

initialize $z_{nd}$ randomly $\forall d \in 1..D$, $\forall n \in 1..N_d$;

compute initial counts $c^k_d$, $\tilde{c}_m^k$, $\tilde{c}^k$ $\forall d \in 1..D$, $\forall k \in 1..K$, $\forall m \in 1..M$;

for $i \leftarrow 1$ to $I$ do
    for $d \leftarrow 1$ to $D$ do
        for $n \leftarrow 1$ to $N_d$ do
            $c^{z_{nd}}_d$--; $\tilde{c}^{z_{nd}}_{w_{nd}}$--; $\tilde{c}^{z_{nd}}$--;
        for $k \leftarrow 1$ to $K$ do
            $p[k] = \frac{\alpha+c^k_d}{K\alpha+N_d-1} \frac{\gamma+\tilde{c}^k_{w_{nd}}}{M\gamma+\tilde{c}^k}$;
        end
        sample $k$ from probability distribution $p[k]$;
        $z_{nd} \leftarrow k$;
        $c^k_d$++; $\tilde{c}^k_{w_{nd}}$++; $\tilde{c}^k$++;
    end
end
Entropy of text

- **joint probability** \( p(\text{TEXT}) = \prod_{i=1}^{N} p(w_i) = \prod_{m=1}^{M} p(m)^{c_m} \)

- **log probability** \( \log p(\text{TEXT}) = \log \prod_{m=1}^{M} p(m)^{c_m} = \sum_{m=1}^{M} c_m \log p(m) \)

- **entropy** \( H(\text{TEXT}) = - \sum_{m=1}^{M} \frac{c_m}{N} \log p(m) = \frac{\log p(\text{TEXT})}{N} \)

- **perplexity** \( 2^H(\text{TEXT}) \)

A perplexity of \( g \) corresponds to the uncertainty associated with a die with \( g \) sides, which generates each new word.
Collapsed sampling for Mixture of Categoricals

\[
p(z_d = k | \{ w \}, \{ z_{-d} \}, \gamma, \alpha) = \\
= \frac{p(z_d = k | \{ z_{-d} \}, \gamma, \alpha) \ p(\{ w \} | z_d = k, \{ z_{-d} \}, \gamma, \alpha)}{p(\{ w \} | \{ z_{-d} \}, \gamma, \alpha)} \\
\propto p(z_d = k | \{ z_{-d} \}, \alpha) \ p(\{ w \} | z_d = k, \{ z_{-d} \}, \gamma) \\
\propto p(z_d = k | \{ z_{-d} \}, \alpha) \ \prod_{n=1}^{N_d} p(w_{nd} | \{ w_{-d} \}, z_d = k, \{ z_{-d} \}, \gamma) \\
\propto \int p(z_d = k | \theta) \ p(\theta | z_{-d}, \alpha) \ d\theta \ \int \prod_{n=1}^{N_d} p(w_{nd} | \beta_k) p(\beta_k | \{ w_{-d} \}, \{ z_{-d} \}, \gamma) \ d\beta_k \\
= \frac{\alpha + c^k - 1}{K \alpha + D - 1} \ \prod_{n=1}^{N_d} \frac{\gamma + \tilde{c}^k_{w_{nd}}}{M \gamma + \sum_{m=1}^{M} \tilde{c}^k_m}
\]

- \( c^k \): How many documents are assigned to topic \( k \).
- \( \tilde{c}^k_m \): How many times the word \( m \) is in a document assigned to topic \( k \).
Mixture of categoricals - algorithm

initialize $z_d$ randomly $\forall d \in 1..D$;
compute initial counts $c^k$, $\tilde{c}_m$, $\tilde{c}^k$ $\forall k \in 1..K$, $\forall m \in 1..M$;
for $i \leftarrow 1$ to $I$ do
  for $d \leftarrow 1$ to $D$ do
    $c^{z_d} \leftarrow$;
    for $n \leftarrow 1$ to $N_d$ do
      $\tilde{c}^{z_d} \leftarrow$; $\tilde{c}^{z_d} \leftarrow$
    end
  for $k \leftarrow 1$ to $K$ do
    $p[k] = \frac{\alpha + c^k}{K\alpha + D - 1} \frac{\gamma + \tilde{c}^k}{M\gamma + \tilde{c}^k}$;
  end
  sample $k$ from probability distribution $p[k]$;
  $z_d \leftarrow k$;
  $c^k \leftarrow$;
  for $n \leftarrow 1$ to $N_d$ do
    $\tilde{c}^{z_d} \leftarrow$; $\tilde{c}^{z_d} \leftarrow$