CSE 291B 2/3

Review

- Document modeling

$$\begin{array}{c|c} \uparrow & \\ V & \begin{bmatrix} \text{Term-by-document} \\ \text{matrix } X_{ij} \\ \\ \end{bmatrix} \\ \downarrow & \\ & \longleftarrow M \longrightarrow \end{array}$$

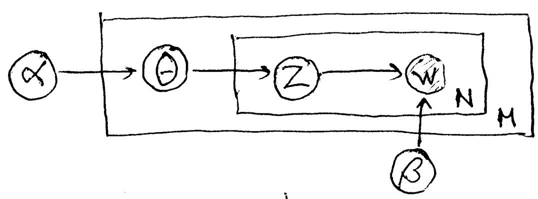
V = # words in vocabulary
M = # documents in corpus
X_{ij} = count of ith word in jth document

- Notation

Document $\vec{w} = (w_1, w_2, \ldots, w_N)$ sequence of N words

Corpus $D = (\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_M)$ collection of M documents

- Latent Dirichlet allocation (LDA)



- Generative model

For each document $\vec{w}$ in corpus D:

1) Choose N = # words
2) Choose topic weights θ from $P(\theta|\alpha) = \dfrac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_k \Theta(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
3) Do N times:
   a) Choose topic $z_n \in \{1, 2, \ldots, k\}$ from $P(z_n|\theta) = \prod_{k=1}^{K} \theta_k^{I(z_n, k)}$ ← indicator function
   b) Choose word $w_n \in \{1, 2, \ldots, V\}$ from $P(w_n = j | z_n = k) = \beta_{kj}$

How to learn parameters $\{\alpha_k\}$ and $\{\beta_{kj}\}$ from data X?

Learning from "complete data" (warm-up)

- Joint probabilities

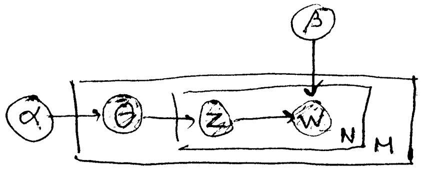
$$P(\theta, \vec{z}, \vec{w} \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \beta)$$

topic weights — topic sequence — word sequence — parameters

(per document)

$$P(\{\theta_m\}_{m=1}^{M}, \{\vec{z}_m\}, D \mid \alpha, \beta) = \prod_{m=1}^{M} P(\theta_m, \vec{z}_m, \vec{w}_m \mid \alpha, \beta)$$

(over corpus)



- Log-likelihood

$$\mathcal{L}(\alpha, \beta) = \underbrace{\sum_m \log P(\theta_m \mid \alpha)}_{\mathcal{L}(\alpha)} + \sum_{mn} \log P(z_{nm} \mid \theta_m) + \underbrace{\sum_{nm} \log P(w_{nm} \mid z_{nm}, \beta)}_{\mathcal{L}(\beta)}$$

- Maximum likelihood (ML) estimation

For multinomial parameters:

$$\beta^* = \underset{\beta}{\operatorname{argmax}}\ \mathcal{L}(\beta)$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0 \longrightarrow \beta^*_{kj} = \frac{\sum_{n,m} I(w_{nm}, j)\, I(z_{nm}, k)}{\sum_{n,m} I(z_{nm}, k)}$$

simple ratio of counts

For dirichlet parameters:

$$\alpha^* = \underset{\alpha}{\operatorname{argmax}}\ \mathcal{L}(\alpha)$$

observed topic weights

$$\mathcal{L}(\alpha) = \sum_m \left\{ \log \Gamma\left(\sum_k \alpha_k\right) - \sum_k \log \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \log \theta_{km} \right\}$$

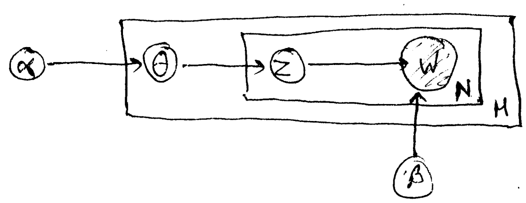
not obvious, but concave α

Define: $\Psi(x) = \frac{d}{dx}[\log \Gamma(x)]$ "digamma function"

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \longrightarrow \frac{1}{M}\sum_{m=1}^{M} \log \theta_{km} = \Psi(\alpha_k) - \Psi\left(\sum_{l=1}^{k} \alpha_l\right)$$

These nonlinear equations have a unique solution for α^*, which can be computed by (say) Newton's method.

$\boxed{\text{Incomplete data}}$



- Inference

$$P(\theta, \vec{z} \mid \vec{w}, \alpha, \beta) = \frac{P(\theta, \vec{z}, \vec{w} \mid \alpha, \beta)}{P(\vec{w} \mid \alpha, \beta)}$$

How to compute denominator?

- Marginal probability of document

$$P(\vec{w} \mid \alpha, \beta) = \int d\theta \sum_{\vec{z}} P(\theta, \vec{z}, \vec{w} \mid \alpha, \beta)$$

product over N words in document

$$= \int d\theta \, P(\theta \mid \alpha) \left\{ \prod_{n=1}^{N} \left[\sum_{z_n=1}^{K} P(z_n \mid \theta) P(w_n \mid z_n, \beta) \right] \right\}$$

$$P(\vec{w}_j \mid \alpha, \beta) = \int_0^1 d\theta_1 \int_0^{1-\theta_1} d\theta_2 \cdots \int_0^{1-\theta_1-\cdots-\theta_{k-1}} d\theta_k \; \frac{\Gamma(\Sigma \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \left\{ \prod_{i=1}^{V} \left(\sum_{k=1}^{K} \theta_k \beta_{kj} \right)^{X_{ij}} \right\}$$

jth document

product over V words in vocabulary, weighted by word-document count X_{ij}

- Log-likelihood of corpus

$$\log P(D \mid \alpha, \beta) = \sum_{j=1}^{M} \log P(\vec{w}_j \mid \alpha, \beta)$$

How to maximize what we cannot compute?

- EM algorithm
  - E-step compute statistics of posterior distribution:
    e.g. $P(z_n=k|\vec{w},\alpha,\beta)$, $E[\log\theta_k|\vec{w},\alpha,\beta]$
  - M-step updates α,β based on these statistics.
  - common approximations for intractable E-step:
    (1) Markov chain Monte Carlo (MCMC)
    (2) Variational method

- Review of MCMC

  To estimate $P(\theta|\vec{w},\alpha,\beta)$ and $P(z_n|\vec{w},\alpha,\beta)$:
  - Fix words $w_1, w_2, \ldots, w_n$ to observed values
  - Initialize hidden nodes $\{\theta_k\}_{k=1}^K$, $\{z_n\}_{n=1}^N$ at random values
    (constrained to sum to 1)
  - Repeat S times:
    - pick hidden node $X \in \{\theta, z_1, \ldots, z_N\}$ at random
    - use Bayes rule to compute $P(X|$ all other nodes at current values$)$
    - resample X from this distribution

- Notation

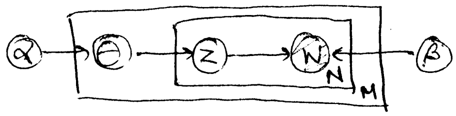
  denote hidden configurations as

$$\{\theta^{(s)}, z_1^{(s)}, \ldots, z_N^{(s)}\}_{s=1}^S$$

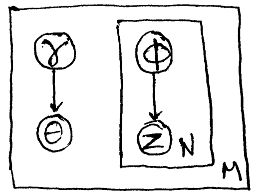
- Estimate $P(z_n=k|\vec{w},\alpha,\beta) = \frac{1}{S}\sum_{s=1}^S I(z_n^{(s)},k)$
- Estimate $P(\theta|\vec{w},\alpha,\beta) = \frac{1}{S}\sum_{s=1}^S \delta(\theta-\theta^{(s)})$

- Variational method

  Approximate intractible $P(\theta, \vec{z} \mid \vec{w}, \alpha, \beta)$ by a tractable distribution $Q(\theta, z \mid \gamma, \vec{\phi})$



$P(\theta, \vec{z} \mid \vec{w}, \alpha, \beta)$



$Q(\theta, z \mid \gamma, \phi)$

if θ and z are peaked, the approx may have the same mode as the original

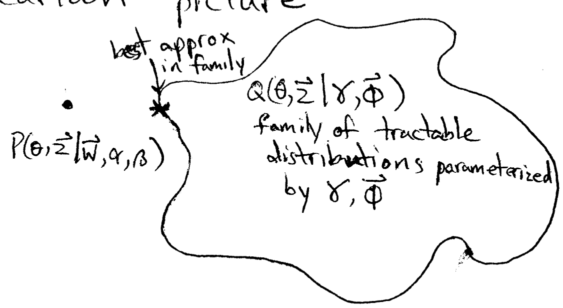
- Explicit form of approximation:

$$Q(\theta, \vec{z} \mid \gamma, \vec{\phi}) = \underbrace{\frac{\Gamma(\sum_k \gamma_k)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1}}_{\text{dirichlet dist.}} \underbrace{\prod_{n=1}^{N} \prod_{k=1}^{K} \phi_{kn}^{I(z_n, k)}}_{\text{distribution over the } K \text{ topics over } N \text{ words}}$$

- Variational inference
  - Choose "variational" parameters $\gamma, \vec{\phi}$ to obtain best approximation:

$$\underset{\gamma, \vec{\phi}}{\operatorname{argmin}} \left[KL\left(Q(\theta, \vec{z} \mid \gamma, \vec{\phi}), P(\theta, \vec{z} \mid \vec{w}, \alpha, \beta)\right) \right]$$

- Cartoon picture



- tractable to compute $KL(Q, P)$ up to constant independent of $\{\gamma, \vec{\phi}\}$

- Useful lower bound

  shorthand: drop dependence on $\alpha, \beta, \gamma, \vec{\Phi}$

$$\log P(\vec{w}) = \log\left[\frac{P(\vec{w},\theta,\vec{z})}{P(\theta,\vec{z}\,|\,\vec{w})}\right] \quad \text{for all } \theta, \vec{z}$$

$$= \int d\theta \sum_{\vec{z}} Q(\theta,\vec{z}) \log\left[\frac{P(\vec{w},\theta,\vec{z})}{P(\theta,\vec{z}\,|\,\vec{w})}\right] \quad \text{for any distribution } Q$$

$$= \int d\theta \sum_{\vec{z}} Q(\theta,\vec{z}) \log\left[\frac{P(\vec{w},\vec{\theta},\vec{z})}{Q(\theta,\vec{w}\,|\,\vec{w})}\right] + \underbrace{KL\left(Q(\theta,\vec{z}),\, P(\theta,\vec{z}\,|\,\vec{w})\right)}_{\geq 0}$$

$$\geq \int d\theta \sum_{\vec{z}} Q(\theta,\vec{z}) \log\left[\frac{P(\vec{w},\theta,\vec{z})}{Q(\theta,\vec{z})}\right]$$