1 Exact Inference

TODO

2 Variational Inference

To be consistent with the LDA paper, \( \nu \) is used to denote the hyperparameter of prior over \( \omega \).

1. Latent variables: \( h = \{ \omega, \theta, \delta, z \} \) for each document, where \( \delta = \{ \delta_i \}_{i=2}^N \) and \( z = \{ z_i \}_{i=1}^N \).
   
   Model parameters: \( u = \{ \nu, \alpha, \beta, A \} \), where \( \beta = \{ \beta_k \}_{k=1}^K \).

2. For each document, the mean-field variational distribution can be defined as

\[
q(\omega, \theta, \delta, z; \eta, \gamma, \lambda, \phi) = q(\omega; \eta) q(\theta; \gamma) \prod_{i=2}^N q(\delta_i; \lambda_i) \prod_{i=1}^N q(z_i; \phi_i), \tag{1}
\]

where each of the terms represents Beta, Dirichlet, Bernoulli, and multinomial density respectively.

3. Complete-data likelihood for each document:

\[
p(h, w; u) = p(\omega; \nu) p(\theta; \alpha) \left[ p(z_i|\theta) \prod_{i=2}^N p(\delta_i|\omega) p(z_i|\delta_i, \theta, z_{i-1}; A) \right] \prod_{i=1}^N p(w_i|z_i; \beta), \tag{2}
\]

where

\[
p(\omega; \nu) = \text{Beta}(\omega; \nu) \tag{3}
\]
\[
p(\theta; \alpha) = \text{Dir}(\theta; \alpha) \tag{4}
\]
\[
p(z_i|\theta) = \text{Mult}(z_i; \theta) \tag{5}
\]
\[
p(\delta_i|\omega) = \text{Ber}(\delta_i; \omega) \tag{6}
\]
\[
p(z_i|\delta_i, \theta, z_{i-1}; A) = p(z_i|\theta)^{\delta_i} p(z_i|z_{i-1}; A)^{1-\delta_i} \tag{7}
\]
\[
p(z_i|z_{i-1}; A) = \text{Mult}(z_i; A_{z_i}) \tag{8}
\]
\[
p(w_i|z_i; \beta) = \text{Mult}(w_i; \beta_{z_i}). \tag{9}
\]

The full joint distribution is the product over all the documents, due to the iid assumption.

4. Evidence lower bound for each document:

\[
\log p(w; u) = \log \int_\Omega \int_\Theta \sum_{\delta} \sum_z p(\omega, \theta, \delta, z, w; u) \, d\theta \, d\omega \tag{10}
\]
\[
= \log \int_\Omega \int_\Theta \sum_{\delta} \sum_z \frac{p(\omega, \theta, \delta, z, w; u)}{q(\omega, \theta, \delta, z)} q(\omega, \theta, \delta, z) \, d\theta \, d\omega \tag{11}
\]
\[
\geq \mathbb{E}_q [\log p(\omega, \theta, \delta, z, w; u)] - \mathbb{E}_q [\log q(\omega, \theta, \delta, z)] \tag{12}
\]
\[
\equiv L(\eta, \gamma, \lambda, \phi). \tag{13}
\]
Expand the first term of the ELBO:

\[
\mathbb{E}_q [\log p(\omega, \theta, \delta, z, w; u)] = \mathbb{E}_q [\log p(\omega; \nu)] + \mathbb{E}_q [\log p(\theta; \alpha)] + \mathbb{E}_q [\log p(z_1; \theta)] \\
+ \sum_{i=2}^N \mathbb{E}_q [\log p(\delta_i; \omega)] + \mathbb{E}_q [\delta_i \log p(z_i | \theta)] + \mathbb{E}_q [(1 - \delta_i) \log p(z_i | z_{i-1}; A)] \\
+ \sum_{i=1}^N \mathbb{E}_q [\log p(w_i | z_i; \beta)],
\]

where

\[
\mathbb{E}_q [\log p(\omega; \nu)] = \log \Gamma(1^T \nu) - 1^T \log \Gamma(\nu) + (\nu - 1)^T (\Psi(\eta) - \Psi(1^T \eta))
\]

\[
\mathbb{E}_q [\log p(\theta; \alpha)] = \log \Gamma(1^T \alpha) - 1^T \log \Gamma(\alpha) + (\alpha - 1)^T (\Psi(\gamma) - \Psi(1^T \gamma))
\]

\[
\mathbb{E}_q [\log p(z_1; \theta)] = \phi_\theta^T (\Psi(\gamma) - \Psi(1^T \gamma))
\]

\[
\mathbb{E}_q [\log p(\delta_i; \omega)] = \lambda_i \phi_\omega^T (\Psi(\eta) - \Psi(1^T \eta))
\]

\[
\mathbb{E}_q [\delta_i \log p(z_i | \theta)] = \lambda_i \phi_i^T (\log \beta) w_i,
\]

\[
\mathbb{E}_q [(1 - \delta_i) \log p(z_i | z_{i-1}; A)] = (1 - \lambda_i) \phi_{i-1}^T (\log A) \phi_i
\]

\[
\mathbb{E}_q [\log p(w_i | z_i; \beta)] = \phi_i^T (\log \beta) w_i
\]

and \(\lambda_i = (\lambda_i, 1 - \lambda_i)^T\), \(z_i\) and \(w_i\) are written as binary vectors with only one entry being nonzero. The derivation is similar to LDA, except for the fact that if \(X \sim \text{Beta}(\alpha, \beta)\), then \(\mathbb{E} [\log (1 - X)] = \Psi(\beta) - \Psi(\alpha + \beta)\).

Expand the second term of the ELBO:

\[
\mathbb{E}_q [\log q(\omega, \theta, \delta, z)] = \mathbb{E}_q [\log q(\omega)] + \mathbb{E}_q [\log q(\theta)] + \sum_{i=2}^N \mathbb{E}_q [\log q(\delta_i)] + \sum_{i=1}^N \mathbb{E}_q [\log q(z_i)],
\]

where

\[
\mathbb{E}_q [\log q(\omega)] = \log \Gamma(1^T \eta) - 1^T \log \Gamma(\eta) + (\eta - 1)^T (\Psi(\eta) - \Psi(1^T \eta))
\]

\[
\mathbb{E}_q [\log q(\theta)] = \log \Gamma(1^T \gamma) - 1^T \log \Gamma(\gamma) + (\gamma - 1)^T (\Psi(\gamma) - \Psi(1^T \gamma))
\]

\[
\mathbb{E}_q [\log q(\delta_i)] = \lambda_i \log \lambda_i
\]

\[
\mathbb{E}_q [\log q(z_i)] = \phi_i^T \log \phi_i
\]

The full ELBO (13) is the difference between (14) and (24).

5. Again this is the updates for a single document.

For \(\eta\): for all \(j = \{1, 2\},\)

\[
\frac{\partial \ell}{\partial \eta_j} = \Psi'(\eta_j) \left( \nu_j + \sum_{i=2}^N \lambda_{ij} - \eta_j \right) + \Psi'(1^T \eta) \cdot 1^T \left( \nu + \sum_{i=2}^N \lambda_i - \eta \right) = 0
\]

\[
\Rightarrow \quad \eta^* = \nu + \sum_{i=2}^N \lambda_i.
\]
For \( \gamma \): for all \( k = \{1, \ldots, K\} \),
\[
\frac{\partial L}{\partial \gamma_k} = \Psi'(\gamma_k) \left( \alpha_k + \sum_{i=1}^{N} \hat{\lambda}_i \phi_{ik} - \gamma_k \right) - \Psi' \left( \mathbf{1}^\top \gamma \right) \cdot \mathbf{1}^\top \left( \alpha + \sum_{i=1}^{N} \hat{\lambda}_i \phi_i - \gamma \right) = 0 \tag{31}
\]
\[
\Rightarrow \gamma^* = \alpha + \sum_{i=1}^{N} \hat{\lambda}_i \phi_i, \tag{32}
\]
where \( \hat{\lambda} = (1, \lambda_2, \ldots, \lambda_N)^\top \).

For \( \lambda_i \):
\[
\frac{\partial L}{\partial \lambda_i} = \Psi(\eta_1) - \Psi(\eta_2) + \phi_i^\top (\Psi(\gamma) - \Psi(\mathbf{1}^\top \gamma)) - \phi_{i-1}^\top \log A \phi_i - \lambda_i^{-1} = 0 \tag{33}
\]
\[
\Rightarrow \lambda_i^* = \left[ \Psi(\eta_1) - \Psi(\eta_2) + \phi_i^\top (\Psi(\gamma) - \Psi(\mathbf{1}^\top \gamma)) - \phi_{i-1}^\top \log A \phi_i \right]^{-1}. \tag{34}
\]

For \( \phi_i \): Since \( \phi \) is a multinomial parameter, it requires additional simplex constraint, whose dual variable is denoted as \( s \). Again, we consider \( k = \{1, \ldots, K\} \),

If \( i = 1 \),
\[
\frac{\partial L}{\partial \phi_{1k}} = \Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma) + (1 - \lambda_{i+1}) \log A_{k:, \phi_{i+1}} + \log \beta_{k2} - \log \phi_{ik} - 1 + s = 0 \tag{35}
\]
\[
\Rightarrow \phi_{1k}^* \propto \beta_{k2} \exp \left\{ \Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma) + (1 - \lambda_{i+1}) \log A_{k:, \phi_{i+1}} \right\}. \tag{36}
\]

If \( i = 2, \ldots, N - 1 \),
\[
\frac{\partial L}{\partial \phi_{ik}} = \lambda_i (\Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma)) + (1 - \lambda_i) \log A_{k:, \phi_{i-1}} + (1 - \lambda_{i+1}) \log A_{k:, \phi_{i+1}}
\]
\[
+ \log \beta_{ki} - \log \phi_{ik} - 1 + s = 0 \tag{38}
\]
\[
\Rightarrow \phi_{ik}^* \propto \beta_{ki} \exp \left\{ \lambda_i (\Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma)) + (1 - \lambda_i) \log A_{k:, \phi_{i-1}} + (1 - \lambda_{i+1}) \log A_{k:, \phi_{i+1}} \right\}. \tag{39}
\]

If \( i = N \),
\[
\frac{\partial L}{\partial \phi_{Nk}} = \lambda_i (\Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma)) + (1 - \lambda_i) \log A_{k:, \phi_{i-1}} + \log \beta_{kN} - \log \phi_{ik} - 1 + s = 0 \tag{40}
\]
\[
\Rightarrow \phi_{Nk}^* \propto \beta_{kN} \exp \left\{ \lambda_i (\Psi(\gamma_k) - \Psi(\mathbf{1}^\top \gamma)) + (1 - \lambda_i) \log A_{k:, \phi_{i-1}} \right\}. \tag{41}
\]

6. For the multinomial and Dirichlet hyperparameters, we can use approximate empirical Bayes using ELBO as a surrogate for the marginal likelihood, as in the LDA paper.

For \( \nu \): Newton’s method with
\[
\frac{\partial L}{\partial \nu_j} = D \left[ \Psi(\mathbf{1}^\top \nu) - \Psi(\nu_j) \right] + \sum_{d=1}^{D} \left[ \Psi(\eta_{dj}) - \Psi(\mathbf{1}^\top \eta_d) \right] \tag{42}
\]
\[
\frac{\partial^2 L}{\partial \nu_j \partial \nu_l} = D \Psi'(\mathbf{1}^\top \nu) - \delta_{j,l} D \Psi'(\nu_j). \tag{43}
\]

For \( \alpha \): Newton’s method with
\[
\frac{\partial L}{\partial \alpha_k} = D \left[ \Psi(\mathbf{1}^\top \alpha) - \Psi(\alpha_k) \right] + \sum_{d=1}^{D} \left[ \Psi(\gamma_{dk}) - \Psi(\mathbf{1}^\top \gamma_d) \right] \tag{44}
\]
\[
\frac{\partial^2 L}{\partial \alpha_k \partial \alpha_l} = D \Psi'(\mathbf{1}^\top \alpha) - \delta_{k,l} D \Psi'(\alpha_k). \tag{45}
\]
For $\beta$: adding a simplex constraint with dual variable $s_k$,

$$\frac{\partial L}{\partial \beta_{kv}} = \sum_{d=1}^{D} \sum_{i=1}^{N} \phi_{dik} \delta_{w_{di},v} \beta_{kv}^{-1} + s_k = 0$$  \hspace{1cm} (46)

$$\Rightarrow \beta_{kv}^* \propto \sum_{d=1}^{D} \sum_{i=1}^{N} \phi_{dik} \delta_{w_{di},v}$$  \hspace{1cm} (47)

For $A$: Optimize subject to the normalization constraint (with dual variable $s_k$),

$$\frac{\partial L}{\partial A_{kl}} = \sum_{d=1}^{D} \sum_{i=2}^{N} (1 - \lambda_{di}) \phi_{d,i-1,k} \phi_{d,i,l} A_{kl}^{-1} + s_k = 0$$  \hspace{1cm} (48)

$$\Rightarrow A_{kl}^* \propto \sum_{d=1}^{D} \sum_{i=2}^{N} (1 - \lambda_{di}) \phi_{d,i-1,k} \phi_{dil}.$$  \hspace{1cm} (49)

3 Markov Chain Monte Carlo

3.1 Sampling Basics

1. We have $U \sim \text{Unif}(0,1)$ and $X \sim F^{-1}(U)$. Evaluate the cdf of $X$ at $t$:

$$\mathbb{P}(X \leq t) = \mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t)) = F(t).$$  \hspace{1cm} (50)

This method is only applicable if the explicit form of $F^{-1}$ is known.

Same result holds for generalized inverse cdf defined as

$$F^{-1}(t) = \inf \{ x : F(x) \geq t \}.$$  \hspace{1cm} (51)

2. We have

$$\int p(x) K_1(x, y) \, dx = p(y) \quad \text{and} \quad \int p(x) K_2(x, y) \, dx = p(y).$$  \hspace{1cm} (52)

Cyclic kernel $K = K_1 \circ K_2$:

$$\int p(x) K(x, z) \, dx = \int p(x) \int K_2(x, y) K_1(y, z) \, dy \, dx$$  \hspace{1cm} (53)

$$= \int K_1(y, z) p(y) \, dy$$  \hspace{1cm} (54)

$$= p(z).$$  \hspace{1cm} (55)

Mixture of kernels $K = \lambda K_1 + (1 - \lambda) K_2$:

$$\int p(x) K(x, y) \, dx = \int p(x) (\lambda K_1(x, y) + (1 - \lambda) K_2(x, y)) \, dx$$  \hspace{1cm} (56)

$$= \lambda p(y) + (1 - \lambda) p(y)$$  \hspace{1cm} (57)

$$= p(y).$$  \hspace{1cm} (58)

Discrete case follows similarly.
3. The transition kernel of Metropolis-Hastings for target density $f(x)$ with proposal $q(y|x)$:

$$K(x, y) = q(y|x)A(x, y) + (1 - r(x)) \delta_x(y), \quad (59)$$

where $r(x) = \int q(y|x)A(x, y) \, dy$ and $\delta_x$ is the Dirac mass at $x$.

We only need to show detailed balance holds:

$$f(x) \left[ q(y|x)A(x, y) + (1 - r(x)) \delta_x(y) \right] = f(y) \left[ q(x|y)A(y, x) + (1 - r(y)) \delta_y(x) \right]. \quad (60)$$

The first term on the LHS:

$$f(x)q(y|x)A(x, y) = f(x)q(y|x) \land f(y)q(x|y). \quad (61)$$

Similarly, the first term on the RHS:

$$f(y)q(x|y)A(y, x) = f(y)q(x|y) \land f(x)q(y|x). \quad (62)$$

Thus (61) and (62) are equal. Also by inspecting the second term, we can notice that they are both zero if $x \neq y$ or are both $1 - r(x)$ if $x = y$.

Thus detailed balance holds for $K$, which means $f(x)$ is the stationary density of the Markov chain.

4. The transition kernel of Gibbs chain is

$$K(x, x') = p_1(x_1'|x_2, \ldots, x_d) \quad (63)
\times p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \times p_d(x_d'|x_1', \ldots, x_{d-1}). \quad (64)$$

Let $p^{(i)}(\cdot)$ be the marginal over all variables except for $x_i$.

$$\int K(x, x')p(x) \, dx = \int p_1(x_1'|x_2, \ldots, x_d) \cdots \quad (65)
\times p^{(1)}(x_2, \ldots, x_d)p(x_1|x_2, \ldots, x_d) \, dx_1 \cdots dx_d \quad (66)
\underbrace{\int K(x, x')p(x) \, dx} = \int p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \quad (67)
\times p(x_1', x_2, \ldots, x_d)p(x_1|x_2, \ldots, x_d) \, dx_1 \cdots dx_d \quad (68)
\text{vanish}
\underbrace{\int K(x, x')p(x) \, dx} = \int p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \quad (69)
\times p^{(2)}(x_1', x_3, \ldots, x_d)p(x_2|x_1', x_3, \ldots, x_d) \, dx_2 \cdots dx_d. \quad (70)$$

Apply recursively, eventually we will arrive at

$$\int K(x, x')p(x) \, dx = p(x_1', \ldots, x_d) = p(x'). \quad (71)$$

Alternatively, we can make use of the fact that each Gibbs sampling subroutine is a special case of Metropolis-Hastings with acceptance rate always equal to 1. However note that the transition kernel (current conditional) is changing within each step, so we need the result on the stationarity of cyclic kernels to complete the proof.
3.2 Gibbs Sampling for SVM

1. Rewrite SVM in the unconstrained form:

\[
w^* = \arg \min_w \frac{\lambda}{2} \|w\|^2_2 + 2 \sum_{i=1}^n \max(0, 1 - y_i w^\top x_i)\]

\[
= \arg \max_w \exp \left\{-\frac{\lambda}{2} \|w\|^2_2\right\} \prod_{i=1}^n \exp \left\{-2 \max(0, 1 - y_i w^\top x_i)\right\}
\]

\[
= \arg \max_w p_0(w) \times \prod_{i=1}^n p(y_i|w, x_i).
\]

Note, however, the prior and likelihood defined this way can be improper.

2. The original posterior can be written as

\[
p(w|y, x) \propto p_0(w) \prod_{i=1}^n \int_0^\infty \frac{1}{\sqrt{2\pi\gamma_i}} \exp \left\{-\frac{1}{2\gamma_i} \left(1 - y_i w^\top x_i + \gamma_i\right)^2\right\} \, d\gamma_i,
\]

which can be regarded as marginalization over \(\gamma\). Hence we can define an augmented posterior distribution as

\[
p(w, \gamma|y, x) \propto p_0(w) \prod_{i=1}^n \frac{1}{\sqrt{2\pi\gamma_i}} \exp \left\{-\frac{1}{2\gamma_i} \left(1 - y_i w^\top x_i + \gamma_i\right)^2\right\}.
\]

3. Let \(\zeta_i = 1 - y_i w^\top x_i\).

\[
p(\gamma_i|\text{rest}) \propto \gamma_i^{-\frac{1}{2}} \exp \left\{-\frac{1}{2} \left(\gamma_i + \frac{\zeta_i^2}{\gamma_i}\right)\right\}
\]

\[
\Rightarrow \gamma_i \sim \text{IG}(1, \zeta_i^2, \frac{1}{2}).
\]

In other words,

\[
\gamma_i^{-1} \sim \text{IG}(|\zeta_i|^{-1}, 1).
\]

4. Let the \(i\)-th column of \(\hat{X}\) be \(y_i x_i\) and let \(\gamma^{-1} = (\gamma_1^{-1}, \ldots, \gamma_d^{-1})^\top\).

\[
p(w|\text{rest}) \propto \exp \left\{-\frac{1}{2} w^\top w - \frac{1}{2} \sum_{i=1}^n \gamma_i^{-1} (1 - y_i w^\top x_i + \gamma_i)^2\right\}
\]

\[
= \exp \left\{-\frac{1}{2} w^\top w - \frac{1}{2} \left(\hat{X}^\top w - (\gamma_1 + 1)^\top \text{diag}(\gamma^{-1}) \left(\hat{X}^\top w - (\gamma_1 + 1)\right)\right)\right\}
\]

\[
\propto \exp \left\{-\frac{1}{2} w^\top \left(\hat{X} \text{diag}(\gamma^{-1}) \hat{X}^\top + \lambda I\right) w + w^\top \left(\hat{X}(1 + \gamma^{-1})\right)\right\}.
\]

Recall the density function for \(z \sim \mathcal{N}(\mu, \Sigma)\) can be written as

\[
p(z) \propto \exp \left\{-\frac{1}{2} z^\top \Sigma^{-1} z + z^\top \Sigma^{-1} \mu\right\}.
\]

Therefore

\[
w \sim \mathcal{N}(\mu, \Sigma),
\]

where \(\Sigma = \left(\hat{X} \text{diag}(\gamma^{-1}) \hat{X}^\top + \lambda I\right)^{-1}\) and \(\mu = \Sigma \hat{X}(1 + \gamma^{-1})\).
5. Given $\Omega$ and $v$, we want to simulate a random vector $r \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma = \Omega^{-1}$ and $\mu = \Sigma v$. We can start by applying Cholesky decomposition to the precision matrix:

$$\Omega = LL^\top.$$  \hfill (86)

Then

$$r = \mu + \Sigma^{\frac{1}{2}} z \quad \Rightarrow \quad L^\top r = L^{-1} v + z.$$  \hfill (87)

Therefore $r$ can be obtained by solving two triangular systems: $r = L^\top \backslash (L \backslash v + z)$. A reference implementation can be found here\footnote{https://github.com/chokkyvista/daSVM/blob/master/mcsvm.m}.

Result is shown in Figure 1. At $T = 600$, the single draw gives 0.9540 accuracy, whereas the averaged samples achieves 0.9690.