Supplementary Material

Joint Discriminative Bayesian Dictionary and Classifier Learning

1 Joint probability distribution

According to the proposed model, the joint probability distribution over the data of the $c^{th}$ class can be expressed as:

$$P(\{y_i^c\}, \{h_i^c\}, \Phi, \Psi, \{z_i^c\}, \{s_i^c\}, \{t_i^c\}, \{\pi_k^c\}, \lambda_y^c, \lambda_t^c, \lambda_y, \lambda_h) =$$

$$\prod_{i=1}^{\left|\mathcal{I}_c\right|} \mathcal{N}(y_i^c|\Phi(z_i^c \odot s_i^c), 1/\lambda_{y_o}I_L) \text{Gam}(\lambda_y|\epsilon_o, f_o) \mathcal{N}(h_i^c|\Psi(z_i^c \odot t_i^c), 1/\lambda_{h_o}I_C) \text{Gam}(\lambda_h|\epsilon_o, f_o)$$

$$\prod_{k=1}^{\left|\mathcal{K}\right|} \mathcal{N}(\varphi_k^c|0, 1/\lambda_{\varphi_o}I_K) \mathcal{N}(\psi_k^c|0, 1/\lambda_{\psi_o}I_C)$$

$$\prod_{i=1}^{\left|\mathcal{I}_c\right|} \prod_{k=1}^{\left|\mathcal{K}\right|} \text{Bernoulli}(z_{ik}^c|\pi_{k_o}^c) \text{Beta}(\pi_k^c a_o, b_o(K - 1))$$

$$\prod_{k=1}^{\left|\mathcal{K}\right|} \mathcal{N}(s_k^c|0, 1/\lambda_{s_o}I_{|\mathcal{K}|}) \text{Gam}(\lambda_s^c|c_o, d_o) \mathcal{N}(t_k^c|0, 1/\lambda_{t_o}I_{|\mathcal{K}|}) \text{Gam}(\lambda_t^c|c_o, d_o).$$

2 Gibbs sampling equations

We have made use of the following theorem [1] while deriving the Gibbs Sampling equations for our model:

**Theorem 1 [1]:** If prior probability over $y_1$ is given as $p(y_1) = \mathcal{N}(y_1|\mu_o, \Lambda_o^{-1})$ and the likelihood function is defined as $p(y_2|y_1) = \mathcal{N}(y_2|Ay_1 + b, L^{-1})$, then the posterior probability distribution over $y_1$ can be written as $p(y_1|y_2) = \mathcal{N}(y_1|\mu, \Lambda^{-1})$, where:

$$\Lambda = \Lambda_o + A^T L A$$

$$\mu = \Lambda^{-1}(A^T L (y - b) + \Lambda_o \mu_o).$$

Below, we derive the sampling equations. The sampling is performed in our approach in an iterative manner. The sampling sequence is the same as the sequence of the equations given below.

**Sample $\varphi_k$:** According to the proposed model, we can write the posterior distribution over the $k^{th}$ dictionary atom $p(\varphi_k|\cdot)$ as follows:

$$p(\varphi_k|\cdot) \propto \prod_{i=1}^{N} \mathcal{N}(y_i|\Phi(z_i \odot s_i), \lambda_{y_o}^{-1}I_L) \mathcal{N}(\varphi_k|0, \lambda_{\varphi_o}^{-1}I_K).$$

We can write the mean of the likelihood function in terms of $\varphi_k$ as:

$$y_{i\varphi_k} = y_i - \Phi(z_i \odot s_i) + \varphi_k(z_{ik} \odot s_{ik}).$$
where \( \mathbf{y}_{i\varphi} \) denotes the contribution of the \( k \)th dictionary atom in approximating \( \mathbf{y}_i \). Hence, the posterior distribution over \( \varphi_k \) can be re-written as:

\[
p(\varphi_k|\cdot) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_{i\varphi_k}|\varphi_k(z_{ik}s_{ik}), \lambda_{y_o}^{-1}\mathbf{I}_L) \mathcal{N}(\varphi_k|0, \lambda_{\varphi_o}^{-1}\mathbf{I}_L).
\]

Exploiting the results of Theorem 1, the posterior over the dictionary atoms can be expressed as:

\[
p(\varphi_k|\cdot) = \mathcal{N}(\varphi_k|\mu_k, \lambda_{\varphi}^{-1}\mathbf{I}_L),
\]

where,

\[
\lambda_{\varphi} = \lambda_{\varphi_o} + \lambda_{y_o} \sum_{i=1}^{N} (z_{ik}s_{ik})^2, \quad \mu_k = \lambda_{y_o} \lambda_{\varphi}^{-1} \sum_{i=1}^{N} (z_{ik}s_{ik}) \mathbf{y}_{i\varphi_k}.
\]

We have arrived at the above expressions by placing \( \mathbf{A} = \sum_{i=1}^{N} (z_{ik}s_{ik}) \) and \( \mathbf{b} = \mathbf{0} \) in the results of Theorem 1. Note that, we have intentionally dropped the super-script ‘c’ from the above expressions. This is because, the dictionary atoms are updated using the training data of all the classes simultaneously. The same is true for updating the columns \( \psi_k \) of the classifier \( \mathbf{\Psi} \).

**Sample \( \psi_k \):** The posterior distribution \( p(\psi_k|\cdot) \) over the \( k \)th column of \( \mathbf{\Psi} \) can be written as:

\[
p(\psi_k|\cdot) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{h}_i|\Psi(z_i \odot t_i), \lambda_{h_o}^{-1}\mathbf{I}_C) \mathcal{N}(\psi_k|0, \lambda_{\psi_o}^{-1}\mathbf{I}_C).
\]

With the same reasoning as for sampling \( \varphi_k \), we can sample \( \psi_k \) from \( p(\psi_k|\cdot) = \mathcal{N}(\psi_k|\mu_k, \lambda_{\psi}^{-1}\mathbf{I}_C) \), where

\[
\lambda_{\psi} = \lambda_{\psi_o} + \lambda_{h_o} \sum_{i=1}^{N} (z_{ik}t_{ik})^2, \quad \mu_k = \lambda_{h_o} \lambda_{\psi}^{-1} \sum_{i=1}^{N} (z_{ik}t_{ik}) \mathbf{h}_{i\psi_k}.
\]

**Sample \( z_{ik} \):** Once the dictionary and the classifier have been sampled, we must sample \( z_{ik} \) based on the updated dictionary and the classifier. The posterior probability distribution over \( z_{ik} \) can be expressed as, \( \forall i \in \mathbf{I}_c, \forall k \in \mathbf{K}: \)

\[
p(z_{ik}^c|\cdot) \propto \mathcal{N}(\mathbf{y}_{i\varphi_k}^c|\varphi_k(z_{ik}^e s_{ik}^e), \lambda_{y_o}^{-1}\mathbf{I}_L) \mathcal{N}(\mathbf{h}_{i\psi_k}^c|\psi_k(z_{ik}^e t_{ik}^e), \lambda_{h_o}^{-1}\mathbf{I}_C) \text{ Bernoulli}(z_{ik}^c|\pi_k^c).
\]

It is straightforward to show that based on the above mentioned posterior

\[
p(z_{ik}^c = 1|\cdot) \propto \pi_{k_o} \exp \left( -\frac{(\mathbf{y}_{i\varphi_k}^c - \varphi_k s_{ik}^e)^\top \lambda_{y_o} \mathbf{I}_L (\mathbf{y}_{i\varphi_k}^c - \varphi_k s_{ik}^e)}{2} \right) \exp \left( -\frac{(\mathbf{h}_{i\psi_k}^c - \psi_k t_{ik}^e)^\top \lambda_{h_o} \mathbf{I}_C (\mathbf{h}_{i\psi_k}^c - \psi_k t_{ik}^e)}{2} \right)
\]

\[
\propto \pi_{k_o} \exp \left( -\frac{\lambda_{y_o} s_{ik}^e^\top s_{ik}^e}{2} \right) \exp \left( -\frac{\lambda_{h_o} (\varphi_k^\top \varphi_k s_{ik}^e s_{ik}^e - 2 s_{ik}^e \mathbf{y}_{i\varphi_k}^c \varphi_k)}{2} \right) 
\]

\[
= \pi_{k_o} \exp \left( -\frac{\lambda_{h_o}}{2} \mathbf{h}_{i\psi_k}^\top \mathbf{h}_{i\psi_k} \right) \exp \left( -\frac{\lambda_{h_o}}{2} (\psi_k^\top \psi_k s_{ik}^e s_{ik}^e - 2 t_{ik}^e \mathbf{h}_{i\psi_k}^\top \psi_k) \right).
\]
Let $p_1 = \pi^c_{k_o} \pi_2 \pi_3 \pi_4$. We can derive an expression for $p(z^c_{ik} = 0|\cdot)$ in a similar fashion, that comes out to be:

$$p(z^c_{ik} = 0|\cdot) \propto (1 - \pi^c_{k_o}) \exp \left( - \frac{\lambda_{y_o} y^c_{i_{\psi_k}}}{2} \right) \cdot \exp \left( - \frac{\lambda_{h_o} h^c_{i_{\psi_k}}}{2} \right).$$

Let $p_o = (1 - \pi^c_{k_o}) \pi_1 \pi_3$. Using $p_1$ and $p_o$, $z^c_{ik}$ can be sampled from the following normalized Bernoulli distribution:

$$z^c_{ik} \sim \text{Bernoulli} \left( \frac{p_1}{p_1 + p_0} \right).$$

Simplifying further:

$$z^c_{ik} \sim \text{Bernoulli} \left( \frac{\pi^c_{k_o} \xi}{1 - \pi^c_{k_o} + \xi \pi^c_{k_o}} \right),$$

where, $\xi = \pi_2 \pi_3$.

**Sample $s^c_{ik}$:** We can write the following regarding the posterior probability distribution over $s^c_{ik}$:

$$p(s^c_{ik}|\cdot) \propto \mathcal{N}(y^c_{i_{\psi_k}}|\varphi_k(z^c_{ik}, s^c_{ik}), \lambda_{y_o} I_L) \mathcal{N}(s^c_{ik}|0, \lambda_{s_o}^{-1}).$$

Exploiting the results of Theorem 1, $s^c_{ik}$ can be sampled from $\mathcal{N}(s^c_{ik}|\mu_s, \lambda_s^{-1})$, where:

$$\lambda_s = \lambda_{s_o} + (\varphi_k z^c_{ik})^\top \lambda_{y_o} I_L (\varphi_k z^c_{ik}) = \lambda_{s_o} + \lambda_{y_o} z^c_{ik} \varphi_k \varphi_k^\top,$$

$$\mu_s = \lambda_s^{-1} (\varphi_k z^c_{ik})^\top \lambda_{y_o} I_L y^c_{i_{\psi_k}} = \lambda_s^{-1} \lambda_{y_o} z^c_{ik} \varphi_k y^c_{i_{\psi_k}}.$$

**Sample $t^c_{ik}$:** Using the same reasoning as for $s^c_{ik}$, we can sample $t^c_{ik}$ from $\mathcal{N}(t^c_{ik}|\mu_t, \lambda_t^{-1})$, where:

$$\lambda_t = \lambda_{t_o} + \lambda_{h_o} z^c_{ik} \psi_k^\top \psi_k,$$

$$\mu_t = \lambda_t^{-1} \lambda_{h_o} z^c_{ik} \psi_k^\top h^c_{i_{\psi_k}}.$$

**Sample $\pi_k$:** We can write the posterior distribution over $\pi_k$ as follows:

$$p(\pi^c_k|\cdot) \propto \prod_{i \in I_c} \text{Bernoulli}(z^c_{ik}|\pi^c_{k_o}) \text{Beta}(\pi^c_{k_o}|a_o/K, b_o(K - 1)/K),$$

$$= \prod_{i \in I_c} \frac{z^c_{ik}}{1 - \pi^c_{k_o}} \times \frac{a_o}{\pi^c_{k_o}} \times \frac{b_o(K - 1)}{K - 1},$$

$$= \prod_{i \in I_c} \frac{\pi^c_{k_o}}{1 - \pi^c_{k_o}} \times \frac{b_o(K - 1)}{K} + |I_c| - \sum_{i = 1}^{|I_c|} z^c_{ik},$$

$$= \text{Beta} \left( \frac{a_o}{K} + \sum_{i = 1}^{|I_c|} \frac{b_o(K - 1)}{K}, b_o(K - 1)/K + |I_c| - \sum_{i = 1}^{|I_c|} z^c_{ik} \right).$$

Thus, we sample $\pi^c_k$ from the above mentioned Beta probability distribution. Note that, in the above derivation we wrote $\pi^c_k$ as $\pi_k$ for readability only.
Sample $\lambda_s^c$: To compute $\lambda_s^c$, we treat $s_{ik}^c$ for all the dictionary atoms simultaneously (we do the same for $\lambda_t^c$ below). We consider $s_{ik}^c \in \mathbb{R}^K$ to be a sample of a Gaussian distribution with isotropic precision. This simplification allows us to efficiently infer the posterior distribution over $\lambda_s^c$ without significantly compromising the performance of our approach. The posterior distribution over $\lambda_s^c$ can be expressed as:

$$
p(\lambda_s^c|\cdot) \propto \prod_{i \in I_c} \mathcal{N}(s_{ik}^c|0, 1/\lambda_s^c_{ik} |K|) \text{Gam}(\lambda_s^c|c_o, d_o)
$$

where $\text{Gam}(\cdot)$ is the well-known gamma function and $\text{det}(\cdot)$ denotes the determinant of a matrix. Neglecting the constants in the right hand side of the above equation, and making use of the property $\text{det}(\lambda I_{|K|}) = \lambda^{|K|}$:

$$
p(\lambda_s^c|\cdot) \propto \lambda_s^c^{(|I_c|\cdot|K|)/2} \exp\left(-\frac{\lambda_s^c}{2} \sum_{i=1}^{|I_c|} s_{ik}^c s_{ik}^c\right) \lambda_s^{c_o - 1} \exp(-d_o \lambda_s^c)
$$

$$
= \lambda_s^{c_o - 1} \exp\left(-\frac{\lambda_s^c}{2} \sum_{i=1}^{|I_c|} s_{ik}^c s_{ik}^c + d_o\right)
$$

$$
\propto \text{Gam}\left(\frac{|I_c||K|}{2} + c_o, \frac{1}{2} \sum_{i=1}^{|I_c|} s_{ik}^c s_{ik}^c + d_o\right).
$$

Therefore, we sample $\lambda_s^c$ as:

$$
\lambda_s^c \sim \text{Gam}\left(\frac{|I_c||K|}{2} + c_o, \frac{1}{2} \sum_{i=1}^{|I_c|} s_{ik}^c s_{ik}^c + d_o\right),
$$

where $\|\cdot\|_2$ denotes the $\ell_2$-norm of a vector.

Sample $\lambda_t^c$: Similarly, we can sample $\lambda_t^c$ from the following Gamma probability distribution:

$$
\lambda_t^c \sim \text{Gam}\left(\frac{|I_t||K|}{2} + c_o, \frac{1}{2} \sum_{i=1}^{|I_t|} t_{ik}^c t_{ik}^c + d_o\right).
$$

Sample $\lambda_y$: The posterior over $\lambda_y$ can be written as:

$$
p(\lambda_y|\cdot) \propto \prod_{i=1}^N \mathcal{N}(y_i|\Phi(z_i \odot s_i), \lambda^{-1}_{y_o} I_L) \text{Gam}(\lambda_y|e_o, f_o).
$$

Again, we have intentionally dropped the superscript 'c' because the computation is performed over the training data of all classes simultaneously. Following similar steps as in the derivations for $\lambda_s^c$ and $\lambda_t^c$ we can show that $\lambda_y$ must be sampled as follows:

$$
\lambda_y \sim \text{Gam}\left(\frac{LN}{2} + e_o, \frac{1}{2} \sum_{i=1}^N \|y_i - \Phi(z_i \odot s_i)\|_2^2 + f_o\right).
$$

Sample $\lambda_h$: Correspondingly, $\lambda_h$ can be sampled as the following:

$$
\lambda_h \sim \text{Gam}\left(\frac{CN}{2} + e_o, \frac{1}{2} \sum_{i=1}^N \|h_i - \Psi(z_i \odot t_i)\|_2^2 + f_o\right).
$$
References