Supplementary Materials: Sparse Representations Optimization with Coupled Bayesian Dictionary and Dictionary Classifier for Efficient Classification

1. Introduction

This document is provided as a supplementary material for the paper titled "Sparse Representations Optimization with Coupled Bayesian Dictionary and Dictionary Classifier for Efficient Classification". We have demonstrated the complete derivation of one of the conditional probabilities of our Bayesian approach as an example. This derivation can be generalized for deriving other probability expressions. However, we have also briefly explained how to derive $p(z_{ik}|-)$

2. Gibbs Sampling

To estimate posterior probabilities in our model, we follow Gibbs sampling inference. We derive the conditional probability of each posterior variable conditioned on other posterior variables and the observed data and use this probability in Gibbs sampling for iteratively drawing samples. The priors used in our approach are in the conjugate exponential family. This facilitates deriving posterior conditional probabilities analytically. The conditional probabilities of posteriors in the following sections have been derived from the overall factorized joint distribution of our model (Figure S1), using the Bayes theorem. The symbol "$|-$" in the following conditional probabilities of the posteriors means conditioned on all variables except the variable of the mentioned probability. Here it is understood that the probability is conditionally independent of all the variables absent in the expression i.e., the variables outside the Markov blanket. This can be inferred from Probabilistic Graphical Model (PGM), Figure S1. The overall joint probability of the model is given below.

$$p(\boldsymbol{\phi}, \mathbf{B}, \mathbf{A}, \mathbf{H}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\lambda}_s, \boldsymbol{\pi}, \lambda_a, \lambda_h) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\phi}_k|\mathbf{0}, \lambda_{\phi_0}^{-1}\mathbf{I}_M) \prod_{k=1}^{K} \mathcal{N}(\mathbf{b}_k|\mathbf{0}, \lambda_{b_0}^{-1}\mathbf{I}_C)$$
$$\prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{a}_{i_{\phi_k}}|\boldsymbol{\phi}_k(z_{ik}.s_{ik}), \lambda_a^{-1}\mathbf{I}_M) \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{h}_{i_{b_k}}|\mathbf{b}_k(z_{ik}.s_{ik}), \lambda_h^{-1}\mathbf{I}_C) \prod_{c=1}^{C}\prod_{i\in I_c}\prod_{k=1}^{K} \text{Bernoulli}(z_{ik}^c|\pi_k^c)$$
$$\prod_{c=1}^{C}\prod_{i\in I_c}\prod_{k=1}^{K} \mathcal{N}(s_{ik}^c|0, (\lambda_s^c)^{-1}) \prod_{c=1}^{C} \text{Gam}(\lambda_s^c|c_o, d_o) \prod_{c=1}^{C}\prod_{k=1}^{K} \text{Beta}\left(\pi_k^c|\tfrac{a_o}{K}, \tfrac{b_o(K-1)}{K}\right)$$
$$\text{Gam}(\lambda_a|e_o, f_o)\text{Gam}(\lambda_h|e_o, f_o)$$

The following example is presented to demonstrate how to derive the conditional probability of a posterior variable, from the overall joint probability of the model, using the Bayes theorem. In the following demonstration, "$|-$" means conditioned on all posterior parameters, and the evidence except $\boldsymbol{\phi}_k$.

Bayes's theorem, in general, is given as $p(\Theta|X) = \frac{p(\Theta \bigcap X)}{p(X)}$

or $p(\Theta|X) \propto p(\Theta \bigcap X) = p(X|\Theta)p(\Theta)$. Let us derive $p(\boldsymbol{\phi}_k|-)$, applying Bayes theorem

i.e., $p(\boldsymbol{\phi}_k|-) = \frac{p(\boldsymbol{\phi},\mathbf{B},\mathbf{A},\mathbf{H},\mathbf{Z},\mathbf{S},\boldsymbol{\lambda}_s,\boldsymbol{\pi},\lambda_y,\lambda_h)}{P(-)}$

Ignoring all expressions that are not dependent upon $\boldsymbol{\phi}_k$, we get

$$P(\boldsymbol{\phi}_k|-) = \frac{\prod_{i=1}^{N} \mathcal{N}(\mathbf{a}_{i_{\phi_k}}|\boldsymbol{\phi}_k(z_{ik}.s_{ik}), \lambda_a^{-1}\mathbf{I}_M)\mathcal{N}(\boldsymbol{\phi}_k|\mathbf{0}, \lambda_{\phi_0}^{-1}\mathbf{I}_M)}{Const} \tag{1}$$

or

$$p(\boldsymbol{\phi}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{a}_{i_{\phi_k}}|\boldsymbol{\phi}_k(z_{ik}.s_{ik}), \lambda_a^{-1}\mathbf{I}_M)\mathcal{N}(\boldsymbol{\phi}_k|\mathbf{0}, \lambda_{\phi_0}^{-1}\mathbf{I}_M) \tag{2}$$

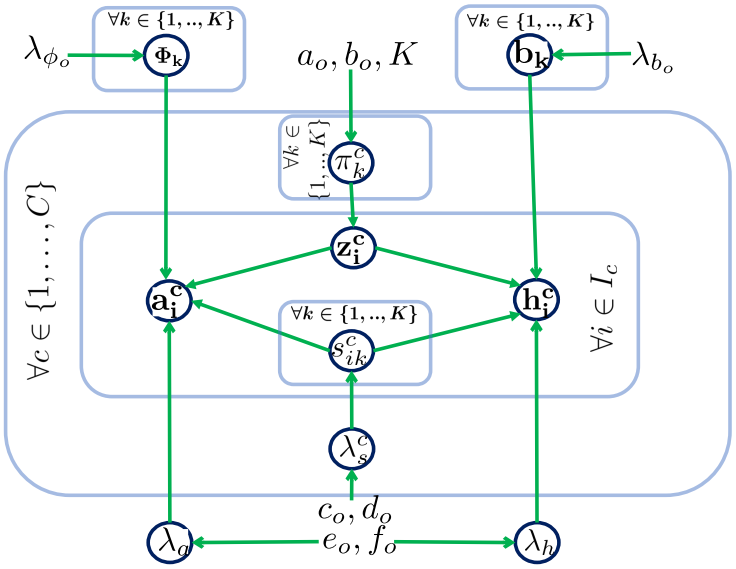

Figure S1. Bayesian Network.

Let $\mathbf{a^j}_{i_{\phi_k}}$ and ${\boldsymbol{\phi}^j}_k$ represents j^{th} components of $\mathbf{a}_{i_{\phi_k}}$ and $\boldsymbol{\phi}_k$, then

$$p({\boldsymbol{\phi}^j}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{a^j}_{i_{\phi_k}}|{\boldsymbol{\phi}^j}_k(z_{ik}.s_{ik}), \lambda_a^{-1}) \mathcal{N}({\boldsymbol{\phi}^j}_k|0, \lambda_{\phi 0}^{-1}) \tag{3}$$

$$p({\boldsymbol{\phi}^j}_k|-) \propto \exp(\sum_{i=1}^{N}(-0.5\lambda_a(\mathbf{a^j}_{i_{\phi_k}} - {\boldsymbol{\phi}^j}_k(z_{ik}.s_{ik}))^2 - 0.5\lambda_{\phi 0}{\boldsymbol{\phi}^j}_k{}^2)) \tag{4}$$

or

$$p({\boldsymbol{\phi}^j}_k|-) \propto \exp(A{\boldsymbol{\phi}^j}_k{}^2 + B{\boldsymbol{\phi}^j}_k) \tag{5}$$

Where
$A = -0.5(\lambda_{\phi_0} + \lambda_a \sum_{i=1}^{N}(z_{ik}.s_{ik})^2)$ and $B = \lambda_a \sum_{i=1}^{N} \mathbf{a^j}_{i_{\phi_k}} z_{ik} s_{ik}$
We can write the above equation as

$$p({\boldsymbol{\phi}^j}_k|-) \propto \exp(A({\boldsymbol{\phi}^j}_k - (-B/2A))^2) \tag{6}$$

or

$$p({\boldsymbol{\phi}^j}_k|-) \propto \exp(-0.5(-2A)({\boldsymbol{\phi}^j}_k - (-B/2A))^2) \tag{7}$$

or

$${\boldsymbol{\phi}^j}_k \sim \mathcal{N}({\boldsymbol{\phi}^j}_k|u, 1/\lambda_\phi) \tag{8}$$

Where
$\lambda_\phi = -2A = \lambda_{\phi_0} + \lambda_a \sum_{i=1}^{N}(z_{ik}.s_{ik})^2$ and $u = B/(-2A) = (1/\lambda_\phi)\lambda_a \sum_{i=1}^{N} \mathbf{a^j}_{i_{\phi_k}} z_{ik} s_{ik}$
Now we can easily convert the uni-variate Gaussian distribution derived above to a multivariate distribution expression. We can sample $\boldsymbol{\phi}_k$ from $\mathcal{N}(\boldsymbol{\phi}_k|\boldsymbol{\mu}_k, \lambda_\phi^{-1}\mathbf{I}_M)$, where
$\lambda_\phi = \lambda_{\phi_o} + \lambda_a \sum_{i=1}^{N}(z_{ik}.s_{ik})^2$, $\boldsymbol{\mu}_k = \lambda_a \lambda_\phi^{-1} \sum_{i=i}^{N}(z_{ik}.s_{ik})\mathbf{a}_{i_{\phi_k}}$

In a similar fashion, the following are the expressions of conditional probabilities of posterior variables derived from the overall joint probability of the model, using Bayes theorem.

Sampling Dictionary Atoms $\boldsymbol{\phi}_k$:

The conditional distribution for taking samples of a dictionary atom may be expressed as

$p(\boldsymbol{\phi}_k|-) \propto \prod_{i=1}^{N} \mathcal{N}(\mathbf{a}_{i_{\phi_k}}|\boldsymbol{\phi}_k(z_{ik}.s_{ik}), \lambda_a^{-1}\mathbf{I}_M)\mathcal{N}(\boldsymbol{\phi}_k|\mathbf{0}, \lambda_{\phi_0}^{-1}\mathbf{I}_M)$

Where, $\mathbf{a}_{i_{\phi_k}} = \mathbf{a}_i - \boldsymbol{\Phi}(\mathbf{z}_i \odot \mathbf{s}_i) + \boldsymbol{\phi}_k(z_{ik} \odot s_{ik})$, is re-construction error induced by all dictionary atoms except k^{th} atom in representing $\mathbf{a}_i$. Here dictionary atom does not carry class label c with it, indicating that we are training a dictionary of the third category where all the atoms are shared for the representation of a data example. $\boldsymbol{\phi}_k$ can be sampled from $\mathcal{N}(\boldsymbol{\phi}_k|\boldsymbol{\mu}_k, \lambda_\phi^{-1}\mathbf{I}_M)$, where

$\lambda_\phi = \lambda_{\phi_o} + \lambda_a \sum_{i=1}^{N} (z_{ik}.s_{ik})^2$, $\boldsymbol{\mu}_k = \lambda_a \lambda_\phi^{-1} \sum_{i=i}^{N} (z_{ik}.s_{ik})\mathbf{a}_{i_{\phi_k}}$

Sampling Classifier Atoms $\mathbf{b}_k$:

Similarly, $\mathbf{b}_k$ can be sampled from $\mathcal{N}(\mathbf{b}_k|\boldsymbol{\mu}_k, \lambda_b^{-1}\mathbf{I}_C)$, where

$\lambda_b = \lambda_{b_o} + \lambda_h \sum_{i=1}^{N} (z_{ik}.s_{ik})^2$, $\boldsymbol{\mu}_k = \lambda_h \lambda_b^{-1} \sum_{i=i}^{N} (z_{ik}.s_{ik})\mathbf{h}_{i_{b_k}}$ Here, $\mathbf{h}_{i_{b_k}}$ is re-construction error induced by all classifier atoms except k^{th} atom in representing $\mathbf{h}_i$. It may be noted here that we use the same weights, $\mathbf{s}_{ik}$, for both the dictionary and the classifier learning.

Sampling z_{ik}^c for assignment of atoms:

The conditional probability for the posterior parameter z_{ik}^c can be expressed as

$p(z_{ik}^c|-) \propto \mathcal{N}(\mathbf{a}_{i_{\phi_k}}^c|\boldsymbol{\phi}_k(z_{ik}^c.s_{ik}^c), \lambda_a^{-1}\mathbf{I}_M)\mathcal{N}(\mathbf{h}_{i_{b_k}}^c|\mathbf{b}_k(z_{ik}^c.s_{ik}^c), \lambda_h^{-1}\mathbf{I}_C)\text{Bernoulli}(z_{ik}^c|\pi_k^c)$.

Let $\mathbf{a^c}^{\mathbf{j}}{}_{i_{\phi_k}}$, $\mathbf{h^c}^{\mathbf{j}}{}_{i_{b_k}}$, $\boldsymbol{\phi}^j{}_k$, and $\boldsymbol{b}^j{}_k$ represents j^{th} components of $\mathbf{a^c}_{i_{\phi_k}}$, $\mathbf{h^c}_{i_{b_k}}$, $\boldsymbol{\phi}_k$, and $\boldsymbol{b}_k$, then

$$p(z_{ik}^c|-) \propto (\pi_{ik}^c)^{z_{ik}^c}(1-\pi_{ik}^c)^{(1-z_{ik}^c)} \exp(-0.5\lambda_a \sum_{j=1}^{M} (\mathbf{a^c}^{\mathbf{j}}{}_{i_{\phi_k}} - \boldsymbol{\phi}^j{}_k(z_{ik}^c.s_{ik}^c))^2$$

$$- 0.5\lambda_h \sum_{j=1}^{C} (\mathbf{h^c}^{\mathbf{j}}{}_{i_{\phi_k}} - \mathbf{b}^{\mathbf{j}}{}_k(z_{ik}^c.s_{ik}^c))^2)$$

putting $z_{ik}^c = 0$ and $z_{ik}^c = 1$ alternatively in the above equation, we can calculate the sampling probability of z_{ik}^c as below $p(z_{ik}^c) \propto \frac{p(z_{ik}^c=1)}{p(z_{ik}^c=1)+p(z_{ik}^c=0)}$, or

z_{ik}^c can be sampled from the following:

$z_{ik}^c \sim \text{Bernoulli}(\frac{\pi_k^c \zeta_1 \zeta_2}{1-\pi_k^c+\zeta_1\zeta_2\pi_k^c})$, where

$\zeta_1 = \exp(-\frac{\lambda_a}{2}(\boldsymbol{\phi}_k^T\boldsymbol{\phi}_k {s_{ik}^c}^2 - 2s_{ik}^c(\mathbf{a}_{i_{\phi_k}}^c)^T\boldsymbol{\phi_k}))$ and

$\zeta_2 = \exp(-\frac{\lambda_h}{2}(\mathbf{b}_k^T\mathbf{b}_k {s_{ik}^c}^2 - 2s_{ik}^c(\mathbf{h}_{i_{b_k}}^c)^T\boldsymbol{b_k}))$

Sampling Sparse Weights s_{ik}^c:

The conditional distribution for s_{ik}^c is

$p(s_{ik}^c|-) \propto \mathcal{N}(\mathbf{a}_{i_{\phi_k}}^c|\boldsymbol{\phi_k}(z_{ik}^c.s_{ik}^c), \lambda_a^{-1}\mathbf{I}_M)$

$\mathcal{N}(\mathbf{h}_{i_{b_k}}^c|\mathbf{b}_k(z_{ik}^c.s_{ik}^c), \lambda_h^{-1}\mathbf{I}_C)N(s_{ik}^c|0, 1/\lambda_s^c)$,

The conjugacy relationship makes it possible to derive distribution analytically as given below

$s_{ik}^c \sim \mathcal{N}(s_{ik}^c|\mu_s, \lambda^{-1})$, where:

$\lambda = \lambda_s^c + \lambda_a {z_{ik}^c}^2 \boldsymbol{\phi}_k^T\boldsymbol{\phi}_k + \lambda_h {z_{ik}^c}^2 \mathbf{b}_k^T\mathbf{b}_k$,

$\mu_s = \lambda^{-1}\left(\lambda_a z_{ik}^c \boldsymbol{\phi}_k^T \mathbf{a}_{i_{\phi_k}}^c + \lambda_h z_{ik}^c \mathbf{b}_k^T \mathbf{h}_{i_{b_k}}^c\right)$,

Here the weights, s_{ik}^c, are learned jointly for the representation of both the training examples and the training labels. This behavior of our approach makes it distinct from others.

Sampling atoms selection probabilities and pruning atoms π_k^c:

$$p(\pi_k^c|-)$$
$$\propto \prod_{i\in I_c} \text{Bernoulli}\left(z_{ik}^c|\pi_k^c\right)\text{Beta}\left(\pi_k^c|\tfrac{a_o}{K}, \tfrac{b_o(K-1)}{K}\right),$$
$$\propto \text{Beta}\left(\tfrac{a_o}{K} + \sum_{i=1}^{|I_c|} z_{ik}^c, \tfrac{b_o(K-1)}{K} + |I_c| - \sum_{i=1}^{|I_c|} z_{ik}^c\right).$$ A dictionary atom ϕ_k is pruned at each iteration of Gibbs sampling according to whether $\sum_{c=1}^{C} \boldsymbol{\pi}_k^c \to 0$ or not. Likewise, classifier atom $\mathbf{b}_k$ is also pruned.

Sampling of Precision parameters for Weights λ_s^c:

$$p(\lambda_s^c|-) \propto \prod_{i\in I_c} \mathcal{N}\left(\mathbf{s}_i^c|\mathbf{0}, 1/\lambda_s^c\mathbf{I}_K\right)\text{Gam}(\lambda_s^c|c_o, d_o).$$
$$\lambda_s^c \sim \text{Gam}\left(\tfrac{|I_c|K}{2} + c_o, \tfrac{1}{2}\sum_{i=1}^{|I_c|}\|s_i^c\|_2^2 + d_o\right).$$

Sampling of Precision Parameter for Data λ_a:

$$p(\lambda_a|-) \propto \prod_{i=1}^{N}\mathcal{N}\left(\mathbf{a}_i|\boldsymbol{\Phi}(\mathbf{z}_i \odot \mathbf{s}_i), \lambda_a^{-1}\mathbf{I}_M\right)\text{Gam}(\lambda_a|e_o, f_o)\ \lambda_a$$
$$\sim \text{Gam}\left(\tfrac{MN}{2} + e_o, \tfrac{1}{2}\sum_{i=1}^{N}\|\mathbf{a}_i - \boldsymbol{\Phi}(\mathbf{z}_i \odot \mathbf{s}_i)\|_2^2 + f_o\right)$$

Sampling of Precision Parameter for Labels λ_h:

$$\text{Similarly, } \lambda_h$$
$$\sim \text{Gam}\left(\tfrac{CN}{2} + e_o, \tfrac{1}{2}\sum_{i=1}^{N}\|\mathbf{h_i} - \mathbf{B}(\mathbf{z_i} \odot \mathbf{s_i})\|_2^2 + f_o\right).$$