# Mathematical Issues

## Written by Okito Yamashita,
## Last revised 2011/03/15

In SLR toolbox, there are seven binary classification algorithms implemented. In this document, the probabilistic models of the implemented classifiers are described and then the brief derivation of algorithms is given. The details of deformation of formula is given in the appendix.

## 0. Notations

- Data

$$\{(\mathbf{x_1}, y_1), \cdots, (\mathbf{x}_N, y_N)\} \;\; \mathbf{x}_i \in R^D, y_i \in \{0,1\}$$

$$X = [\mathbf{x_1}, \cdots, \mathbf{x}_N], \; \mathbf{y} = [y_1, \cdots, y_N]$$

- Linear discriminant function

$$f(\mathbf{x}; \mathbf{w}) = \sum_{d=1}^{D} w_d x_d + w_0 = \mathbf{w}^t \mathbf{x}$$

where $\mathbf{w} = [w_0, w_1, \cdots, w_D], \; \mathbf{x} \to [1, x_1, \cdots, x_D]$

- Sigmoid function (Logistic function)

$$\sigma(x) = 1 / \left(1 + \exp(-x)\right)$$

- Normal distribution

$$N(\mathbf{x}; \mu, S) = 2\pi^{-\frac{d}{2}} \,|\, S \,|^{-\frac{1}{2}} \exp\left( -\frac{1}{2} (\mathbf{x} - \mu)^t S^{-1} (\mathbf{x} - \mu) \right)$$

$$E(\mathbf{x}) = \mu, \quad V(\mathbf{x}) = S$$

where E(x) and V(x) represent expectation and variance of a random variable x, respectively.

- Gamma distribution

$$\Gamma(\alpha; \alpha_0, \gamma_0) = \frac{1}{\Gamma(\gamma_0)} \left( \frac{\gamma_0}{\alpha_0} \right)^{\gamma_0} \alpha^{\gamma_0 - 1} \exp\left( -\frac{\gamma_0}{\alpha_0} \alpha \right)$$

$$E(\alpha) = \alpha_0, \quad V(\alpha) = \frac{\alpha_0^2}{\gamma_0}$$

# 1. Probabilistic Model

All the probabilistic models introduced here can be described in the Bayesian model, which has the likelihood function and a prior distribution of weight (or boundary) parameters. All the models share the identical likelihood function known as the logistic regression (Eq.1) while they uses the different prior distributions.

## Likelihood Model (Logistic Regression (LR)

The logistic regression model is a probabilistic model for binary data developed in the field of statistics. The likelihood function of LR is given in the following form,

$$P(\mathbf{y} \mid X, \mathbf{w}) = \prod P(y_n \mid \mathbf{x}_n, \mathbf{w}) = \prod \sigma_n^{y_n} (1 - \sigma_n)^{1-y_n} \qquad (1)$$

where $\sigma_n = \sigma(\mathbf{w}^t \mathbf{x}_n)$ and the linear boundary is assumed.

## 1.1 Sparse Logistic Regression (SLR)

The likelihood function is given by Eq.(1).

The prior distribution of SLR has the following hierarchical form;

$$For \quad d = 1, \cdots, D$$

$$\begin{cases} P_0(w_d \mid \alpha_d) = N(0, \alpha_d^{-1}) \\ \qquad P_0(\alpha_d) = \alpha_d^{-1} \end{cases} \qquad (2)$$

This hierarchical distributions are known as the automatic relevance determination (ARD) priors in the sparse Bayesian learning literature. The parameter $\alpha_d$ is called the relevance parameter that represents relevance of the corresponding weight parameter. The larger this value is, less relevant the corresponding weight parameter is. If the relevance parameter is marginalized beforehand, we know that the hierarchical prior distributions are equivalent to the following prior distribution of weight parameters,

$$P_0(w_d) = 1/\,|\,w_d\,| \qquad d = 1, \cdots, D \ .$$

The hierarchical form makes it easier to calculate the posterior distribution in an analytical way. Thus the equation (2) is used.

## 1.2 Regularized Logistic Regression (RLR)

The likelihood function is given by Eq.(1).

The prior distribution of RLR is the following multivariate Gaussian distribution;

$$P_0(\mathbf{w}) = N(0, \alpha^{-1} I_D) \ . \qquad (3)$$

$I_d$ is the identity matrix of size $D \times D$.

### 1.3 Linear Relevance Vector Machine (RVM)

The likelihood function is given by Eq.(1) except the linear discriminant function being represented by the linear kernel as follows,

$$\begin{aligned} g(\mathbf{x};\mathbf{w}) &= \sum_{n=1}^{N} w_n \mathbf{x}_n^t \mathbf{x} + w_0 \\ &= \mathbf{k}^t(\mathbf{x})\mathbf{w} \end{aligned}$$

where $\mathbf{k}(\mathbf{x}) = (\mathbf{x}_1^t \mathbf{x}, \cdots, \mathbf{x}_n^t \mathbf{x}, 1)^t$ and $\mathbf{w} = (w_1, \cdots, w_n, w_0)^t$.

Thus the likelihood for each single data is represented by $\sigma_n = \sigma(\mathbf{k}(\mathbf{x}_n)\mathbf{w})$.
The prior distribution of RVM is the ARD prior (Eq(2)).

It should be noted that the sparseness of SLR and RVM results in the different interpretation. The parameters in SLR are associated with features, while the parameters in SVM are associated with samples (note that the number of weight parameters in RVM is *N* not *D*). If the parameters in SLR are estimated in a sparse way, it can be interpreted as feature selection process but that of RVM can be interpreted as sample selection process (similar to 'support vectors' in SVM).

Another thing to be noticed is that the extension to the nonlinear discriminant function is easily realized in RVM because RVM uses the kernel representation. Changing the linear kernel to any other nonlinear kernel such as Gaussian kernel, polynomial kernel can model a non-linear boundary.

### 1.4 L1-Sparse Logistic Regression (L1-SLR)

The likelihood function is given by Eq.(1).
The prior distribution of L1-SLR is the following Laplace distribution;

$$P_0(w_d) = \frac{1}{2}\sqrt{\lambda}\exp(-\sqrt{\lambda}\,|\,w_d\,|) \qquad d = 1, \cdots, D \, . \qquad (4)$$

This Laplace prior can be expressed in the hierarchical form as follows;

$$\begin{cases} P_0(w_d \mid \alpha_d) = N(0, \alpha_d) \\ P_0(\alpha_d) = \dfrac{\lambda}{2}\exp\left(-\dfrac{\lambda}{2}\alpha_d\right) \end{cases} \quad d = 1, \cdots, D \tag{5}$$

Note that $\alpha_d$ is variance of Gaussian distribution in Eq.(5) whereas it is precision (inverse variance) of Gaussian distribution in Eq.(2).

## 2. Derivation of Algorithms

Since all the probabilistic models described above are Bayesian models, the task to estimate weight parameters is to calculate the posterior probability distribution of weight parameters given by

$$P(\mathbf{w} \mid \mathbf{y}) = \frac{\int P(\mathbf{y} \mid \mathbf{w}) P_0(\mathbf{w} \mid \boldsymbol{\alpha}) P_0(\boldsymbol{\alpha}) d\boldsymbol{\alpha}}{\int P(\mathbf{y} \mid \mathbf{w}) P_0(\mathbf{w} \mid \boldsymbol{\alpha}) P_0(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}\, d\mathbf{w}}.$$

Here and hereafter the dependency on X is omitted for notational simplicity. Unfortunately the integrals of numerator and denominator are not analytically tractable. Therefore some approximation method should be applied. In the algorithm derivation, we apply the variational Bayesian method (VB) that assumes the conditional independence condition on posterior distributions and then solve the posterior calculation by maximizing a specific criteria (called free energy).

At first VB defines the free energy using the test function Q();

$$FE(Q(\mathbf{w}, \boldsymbol{\alpha})) = \int Q(\mathbf{w}, \boldsymbol{\alpha}) \log \frac{P(\mathbf{y}, \mathbf{w}, \boldsymbol{\alpha})}{Q(\mathbf{w}, \boldsymbol{\alpha})} d\boldsymbol{\alpha} d\mathbf{w}. \tag{6}$$

The notable issue in this equation is that FE is maximized when and only when the test function Q is equal to the joint posterior distribution $P(\mathbf{w}, \boldsymbol{\alpha} \mid \mathbf{y})$. In addition the maximized value is equivalent to the evidence $P(\mathbf{y})$. Therefore maximizing FE with respect to test function corresponds to finding the joint posterior distribution and computing the evidence. However this functional maximization is as difficult as the original problem, thus it can not be solved directly. VB solves this problem by restricting the test function to having some functional form. In our application, the test function assumes to satisfy the conditional independence condition as

$$Q(\mathbf{w}, \boldsymbol{\alpha}) = Q(\mathbf{w}) Q(\boldsymbol{\alpha}). \tag{7}$$

Under this condition, FE is maximized by alternately maximizing FE with respect to $Q(\mathbf{w})$ and $Q(\boldsymbol{\alpha})$. These steps are equivalent to computing the following W-step and A-step (for more mathematical details, see Bishop 2006).

W-step： $\log Q(\mathbf{w}) = < \log P(\mathbf{y}, \mathbf{w}, \boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})}$

A-step： $\log Q(\boldsymbol{\alpha}) = < \log P(\mathbf{y}, \mathbf{w}, \boldsymbol{\alpha}) >_{Q(\mathbf{w})}$

where $< f(\mathbf{x}) >_{Q(\mathbf{x})}$ denotes the expectation of $f(\mathbf{x})$ with respect to a probabilistic measure $Q(\mathbf{x})$. Concrete algorithms to compute the posterior distributions as well as the posterior mean (the estimate of the weights) can be derived by substituting the probabilistic models described in section 1. But unfortunately even with this approximation, W-step is the analytically intractable for the logistic regression model since the prior distribution and the likelihood function are not conjugate. A further approximation to $Q(\mathbf{w})$ is required. There are two approximation methods for this purpose; Laplace approximation and variational approximation. Both of the methods use the Gaussian distribution as the approximate distribution. This is why there are two algorithms in the toolbox even for one probabilistic model SLR (also RLR).

## 2.1. SLR

### 2.1.1. SLR with Laplace approximation (SLR-LAP)

This algorithm uses the Laplace approximation that approximates the posterior distribution with Gaussian distribution around the MAP estimate.

<u>W-step:</u>

$$\log Q(\mathbf{w}) = \sum_{n=1}^{N} \{ y_n \log \sigma_n + (1 - y_n) \log(1 - \sigma_n) \} - \frac{1}{2} \mathbf{w}^t \bar{A} \mathbf{w} + const \qquad (8)$$

where $\bar{A} = diag(\bar{\alpha}_1, \cdots, \bar{\alpha}_D)$. Let's denote the right-hand side of Eq (8) by $E(\mathbf{w})$. $E(\mathbf{w})$ is not the quadratic form of w, thus $Q(\mathbf{w})$ is not the Gaussian distribution and analytically intractable. But if we notice that $E(\mathbf{w})$ is the exactly same form as the log-likelihood function of logistic regression with the regularization term, this function is well approximated by the quadratic function at the maximizer $\bar{\mathbf{w}}$ (Laplace approximation). This maximization is done by the Newton-Rapson method using the following gradient and Hessian

$$\frac{\partial E}{\partial \mathbf{w}} = X\boldsymbol{\delta} - \bar{A}\mathbf{w} \,,$$

$$\frac{\partial^2 E}{\partial \mathbf{w} \partial \mathbf{w}^t} = -XBX' - \bar{A} \equiv -H(\mathbf{w})$$

where $\boldsymbol{\delta}, X, B$ are given by

$$\begin{aligned}
&\boldsymbol{\delta} = [y_1 - \sigma_1, \cdots, y_N - \sigma_N]^t \;:\; N \times 1 \\
&X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \;:\; D \times N \\
&B = diag(\sigma_1(1-\sigma_1), \cdots, \sigma_N(1-\sigma_N)) \;:\; N \times N
\end{aligned} .$$

Then we obtain the quadratic approximation of $\log Q(\mathbf{w})$ (or equivalently $E(\mathbf{w})$) around the maximizer $\bar{\mathbf{w}}$,

$$\log Q(\mathbf{w}) \approx E(\bar{\mathbf{w}}) - \frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^t H(\bar{\mathbf{w}})(\mathbf{w} - \bar{\mathbf{w}}) . \qquad (9)$$

Thus $Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S)$ *where* $S \equiv H(\bar{\mathbf{w}})^{-1}$.

<u>A-step:</u>

$$\log Q(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{d=1}^{D}\left(\alpha_d(\bar{w}_d^2 + s_d^2) + \log \alpha_d\right) + const$$

Thus $Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d)$ where $\bar{\alpha}_d = \dfrac{1}{\bar{w}_d^2 + s_d^2}$, $\bar{\gamma}_d = \dfrac{1}{2}$ (10)

$\bar{w}_d, s_d^2$ are the *d*th element and the *d*th diagonal element of $\bar{\mathbf{w}}$ and $S$, respectively.

This updating rule is very slow to converge. Instead we use the following fast updating rule motivated by Mackay's effective number of parameters.
Fast updating rule:

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d) \text{ where } \bar{\alpha}_d = \frac{1 - \bar{\alpha}_d s_d^2}{\bar{w}_d^2},\ \bar{\gamma}_d = \frac{1}{2}. \quad (11)$$

Note that the numerator $1 - \bar{\alpha}_d s_d^2$ of Eq.(11) is related to Mackay's effective number of parameters (see section 3.5.3 of Bishop 2006). If this value is close to 1, the corresponding parameter $\bar{w}_d$ can be mainly determined by observations, whereas if this value is close to 0, the corresponding parameter is not sensitive to observations and determined by prior information.

### 2.1.2. SLR with variational approximation (SLR-VAR)

This algorithm first approximate the logistic function with a Gaussian distribution with one auxiliary variable (variational approximation). According to Jaakkola and Jordan 2000, the logistic function can be lower-bounded by

$$\sigma(x) \geq \sigma(\xi)\exp\left\{\frac{x-\xi}{2} - \lambda(\xi)(x^2 - \xi^2)\right\} \text{ where } \lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right) > 0 \ (\xi > 0) .$$

Thus the likelihood function is bounded by

$$P(\mathbf{y} \mid \mathbf{w}) \ge \prod \sigma(\xi_n) \exp\left\{ \frac{z_n - \xi_n}{2} - \lambda(\xi_n)(z_n^2 - \xi_n^2) - z_n(1 - y_n) \right\} \equiv h(\mathbf{y}, \mathbf{w}, \boldsymbol{\xi})$$

with $z_n = \mathbf{x}_n^t \mathbf{w}$. Using this formula, FE is lower-bounded as follows,

$$FE(Q(\mathbf{w},\boldsymbol{\alpha})) \ge \int Q(\mathbf{w},\boldsymbol{\alpha}) \log \frac{h(\mathbf{y},\mathbf{w},\boldsymbol{\xi}) P_0(\mathbf{w} \mid \boldsymbol{\alpha}) P_0(\boldsymbol{\alpha})}{Q(\mathbf{w},\boldsymbol{\alpha})} d\boldsymbol{\alpha} d\mathbf{w} \equiv FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi}).$$

Thus maximizing $FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi})$ with respect to $Q(\mathbf{w}), Q(\boldsymbol{\alpha}), \boldsymbol{\xi}$ alternately gives us the posterior distributions and optimal variational parameters $\boldsymbol{\xi}$. Since the function $h(\mathbf{y},\mathbf{w},\boldsymbol{\xi})$ is the quadratic function of $\mathbf{w}$ for fixed $\boldsymbol{\xi}$, $Q(\mathbf{w})$ becomes the Gaussian distribution, which is easy to calculate analytically. See Bishop and Tipping 2000 for more details

<u>W-step</u>:

$$\log Q(\mathbf{w}) = -\frac{1}{2} \mathbf{w}^t \left( \bar{A} + \sum_{n=1}^{N} 2\lambda(\xi_n) \mathbf{x}_n \mathbf{x}_n^t \right) \mathbf{w} + \left( \sum_{n=1}^{N} (y_n - 0.5) \mathbf{x}_n^t \right) \mathbf{w} + const$$

$$= -\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^t S^{-1} (\mathbf{w} - \bar{\mathbf{w}}) + const$$

Thus $Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S)$ where

$$\begin{cases} S^{-1} = \bar{A} + \sum_{n=1}^{N} 2\lambda(\xi_n) \mathbf{x}_n \mathbf{x}_n^t = \bar{A} + 2X\Lambda X^t \ : D \times D \\ \bar{\mathbf{w}} = S \sum_{n=1}^{N} (y_n - 0.5) \mathbf{x}_n = (\bar{A} + 2X\Lambda X^t)^{-1} X \tilde{\mathbf{y}} \quad : D \times 1 \end{cases} \tag{12}$$

$$\tilde{\mathbf{y}} = [y_1 - 0.5, \cdots, y_N - 0.5]^t \ : \ N \times 1$$
$$X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \ : \ D \times N$$
$$\Lambda = diag(\lambda(\xi_1), \cdots, \lambda(\xi_N)) \ : \ N \times N$$

The computation of $\bar{\mathbf{w}}$ requires the inverse of $\bar{A} + 2X\Lambda X^t$, which is heavy when the number of features (=D) is large. The use of Sherman-Morrison-Woodbury formula for matrix inversion reduces this burden.

$$\begin{cases} S^{-1} = \bar{A} + \sum_{n=1}^{N} 2\lambda(\xi_n) \mathbf{x}_n \mathbf{x}_n^t = \bar{A} + 2X\Lambda X^t \ : D \times D \\ \bar{\mathbf{w}} = \bar{A}^{-1} X (\Lambda^{-1} + 2X^t \bar{A}^{-1} X)^{-1} \Lambda^{-1} \tilde{\mathbf{y}} \ : \ D \times 1 \end{cases}$$

Since the size of matrix which requires inversion ($\Lambda^{-1} + 2X^t \bar{A}^{-1} X$) is $N \times N$, this is operated much faster when the number of samples is less than the number of features.

A-step:

A-step is identical to that of SLR-LAP (see Eq.(11)).

Fast updating rule :

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d) \text{ where } \bar{\alpha}_d = \frac{1 - \bar{\alpha}_d s_d^2}{\bar{w}_d^2},\ \bar{\gamma}_d = \frac{1}{2} \qquad (13)$$

$\xi$ -step:

Taking partial derivative of $FE(Q(\mathbf{w}, \boldsymbol{\alpha}), \boldsymbol{\xi})$ and setting it to zero, we have

$$\xi_n^2 = \mathbf{x}_n^t (\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)\mathbf{x}_n \qquad n = 1, \cdots, N \qquad (14)$$

By noticing that $\xi_n^2$ is the *n*th diagonal element of $X^t(\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)X$ , we obtain

$$\xi_n^2 = \left[ X^t \bar{\mathbf{w}}\bar{\mathbf{w}}^t X + X^t \bar{A}^{-1} X \left( (2\Lambda)^{-1} + X^t \bar{A}^{-1} X \right)^{-1} (2\Lambda)^{-1} \right]_{nn}$$

where the right hand side means (*n,n*) element of the matrix. Here the Sherman-Morrison-Woodbury formula was used again.

## 2.2. RLR

### 2.2.1. RLR with Laplace approximation (RLR-LAP)

W-step:

If $\bar{A} = diag(\bar{\alpha}_1, \cdots, \bar{\alpha}_D)$ in SLR-LAP algorithm is replaced with $\bar{A} = \bar{\alpha} I_D$, this step is identical to SLR-LAP algorithm.

A-step:

$$\log Q(\alpha) = -\frac{1}{2} < \mathbf{w}^t \mathbf{w} > \alpha + \left( \frac{D}{2} - 1 \right) \log \alpha + const$$

Thus $Q(\alpha) = \Gamma(\alpha\ ; \bar{\alpha}, \bar{\gamma})$ where $\bar{\alpha} = \dfrac{D}{\sum_{d=1}^{D} \left( \bar{w}_d^2 + s_d^2 \right)},\ \bar{\gamma} = \dfrac{D}{2}$ (15)

Fast updating rule： $Q(\alpha)=\Gamma(\alpha\ ;\bar{\alpha},\bar{\gamma})$ where $\bar{\alpha}=\dfrac{D-\bar{\alpha}\sum_{d=1}^{D}s_d^2}{\sum_{d=1}^{D}\bar{w}_d^2}$, $\bar{\gamma}=\dfrac{D}{2}$.(16)

### 2.2.2. RLR with variational approximation (RLR-VAR)

W-step:

If $\bar{A}=diag(\bar{\alpha}_1,\cdots,\bar{\alpha}_D)$ in SLR-VAR algorithm is replaced with $\bar{A}=\bar{\alpha}I_D$, this step is identical to SLR-VAR algorithm.

A-step:

A-step is equivalent to that of RLR-LAP.

Fast updating rule： $Q(\alpha)=\Gamma(\alpha\ ;\bar{\alpha},\bar{\gamma})$ where $\bar{\alpha}=\dfrac{D-\bar{\alpha}\sum_{d=1}^{D}s_d^2}{\sum_{d=1}^{D}\bar{w}_d^2}$, $\bar{\gamma}=\dfrac{D}{2}$.

$\xi$ -step:

This step is identical to that of SLR-VAR since this step does not depend on the prior distribution.

$$\xi_n^2=\mathbf{x}_n^t(\bar{\mathbf{w}}\bar{\mathbf{w}}^t+S)\mathbf{x}_n \qquad n=1,\cdots,N$$

## 2.3 RVM

The algorithm implemented for RVM is identical to SLR-VAR except the boundary being modeled by the linear kernels. Thus if $z_n=\mathbf{x}_n^t\mathbf{w}$ in SLR-VAR algorithm is replaced with $z_n=\mathbf{k}^t(\mathbf{x}_n)\mathbf{w}$, the derivation is identical. See Tipping 2001 for the original RVM algorithm.

W-step:

$$\log Q(\mathbf{w})=-\frac{1}{2}\mathbf{w}^t\left(\bar{A}+\sum_{n=1}^{N}2\lambda(\xi_n)\mathbf{k}(\mathbf{x}_n)\mathbf{k}^t(\mathbf{x}_n)\right)\mathbf{w}+\left(\sum_{n=1}^{N}(y_n-0.5)\mathbf{k}^t(\mathbf{x}_n)\right)\mathbf{w}+const$$

$$=-\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}})^tS^{-1}(\mathbf{w}-\bar{\mathbf{w}})+const$$

Thus $Q(\mathbf{w})\sim N(\bar{\mathbf{w}},S)$ where

$$\begin{cases} S^{-1} = \bar{A} + \sum_{n=1}^{N} 2\lambda(\xi_n)\mathbf{k}(\mathbf{x}_n)\mathbf{k}^t(\mathbf{x}_n) = \bar{A} + 2X\Lambda X^t : (N+1)\times(N+1) \\ \bar{\mathbf{w}} = S\sum_{n=1}^{N}(y_n - 0.5)\mathbf{k}(\mathbf{x}_n) \quad : \quad (N+1)\times 1 \end{cases} \tag{17}$$

$$X = [\mathbf{k}(\mathbf{x}_1),\cdots,\mathbf{k}(\mathbf{x}_N)] \ : \ (N+1)\times N$$
$$\Lambda = diag(\lambda(\xi_1),\cdots,\lambda(\xi_N)) \ : \ N\times N$$

<u>A-step:</u>

A-step is identical to that of SLR-LAP (see Eq.(11)).

Fast updating rule:

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d) \ \text{ where } \ \bar{\alpha}_d = \frac{1-\bar{\alpha}_d s_d^2}{\bar{w}_d^2}, \ \bar{\gamma}_d = \frac{1}{2} \tag{18}$$

<u>$\xi$ -step:</u>

Taking partial derivative of $FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi})$ and setting it equal to zero, we have

$$\xi_n^2 = \mathbf{k}^t(\mathbf{x}_n)(\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)\mathbf{k}(\mathbf{x}_n) \qquad n = 1,\cdots,N\,. \tag{19}$$

## 2.4 L1-SLR

The algorithm of L1-SLR-LAP is obtained by modifying the derivation in Krishnapuram et.al 2004. The algorithm of L1-SLR-c is based on component-wise update procedure with the surrogate function in Krishnapuram et.al 2005.

### 2.4.1. L1-SLR with Laplace approximation (L1-SLR-LAP)

We use the hierarchical form of the Laplace prior (Eq.5). The derivation is very similar to that of SLR-LAP.

<u>W-step:</u>

$$\log Q(\mathbf{w}) = \sum_{n=1}^{N}\{y_n \log\sigma_n + (1-y_n)\log(1-\sigma_n)\} - \frac{1}{2}\mathbf{w}^t\bar{V}\mathbf{w} + const \tag{20}$$

where $\bar{V} = diag(<\alpha_1^{-1}>,\cdots,<\alpha_D^{-1}>)$ . $Q(\mathbf{w})$ is obtained in the same way (Newton-Rapson method) as W-step of SLR-LAP with $\bar{A}$ replaced with $\bar{V}$ .

A-step:

$$\log Q(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{d=1}^{D}\left( -\frac{\overline{w}_d^2 + s_d^2}{\alpha_d} - \lambda\alpha_d - \log\alpha_d \right) + const$$

Here $\overline{w}_d, s_d^2$ denote the *d*th element and the *d*th diagonal element of the weight posterior distribution.

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d)$$
$$Q(\alpha_d) = C\alpha_d^{-\frac{1}{2}} \exp\left( -\frac{\overline{w}_d^2 + s_d^2}{2\alpha_d} - \frac{\lambda}{2}\alpha_d \right) . \qquad (21)$$

Unfortunately this distribution is not the Gamma distribution unlike SLR model. But only necessary quantity in W-step is the expectation of $\alpha_d^{-1}$ and this value can be analytically computed as follows,

$$\begin{aligned}
<\alpha_d^{-1}> &= \int \alpha_d^{-1} Q(\alpha_d) d\alpha_d \\
&= \frac{\int \alpha_d^{-\frac{3}{2}} \exp\left( -\frac{\overline{w}_d^2 + s_d^2}{2\alpha_d} - \frac{\lambda}{2}\alpha_d \right) d\alpha_d}{\int \alpha_d^{-\frac{1}{2}} \exp\left( -\frac{\overline{w}_d^2 + s_d^2}{2\alpha_d} - \frac{\lambda}{2}\alpha_d \right) d\alpha_d} \\
&= \frac{\sqrt{\frac{2\pi}{\overline{w}_d^2 + s_d^2}} \exp\left( -2\sqrt{\frac{\lambda}{2} \cdot \frac{\overline{w}_d^2 + s_d^2}{2}} \right)}{\sqrt{\frac{2\pi}{\lambda}} \exp\left( -2\sqrt{\frac{\lambda}{2} \cdot \frac{\overline{w}_d^2 + s_d^2}{2}} \right)} \\
&= \frac{\sqrt{\lambda}}{\sqrt{\overline{w}_d^2 + s_d^2}}
\end{aligned} \qquad (22)$$

The formula derived from EM algorithm is very similar (see Krishnapuram et.al 2004),

$$<\alpha_d^{-1}> = \frac{\sqrt{\lambda}}{\sqrt{\overline{w}_d^2}} . \qquad (23)$$

Only the difference is whether the posterior variance of weights is included in the denominator. My little experience showed that Eq.(23) does work well but Eq.(22 does not. Thus Eq.(23) is applied in the current implementation. Reasons for this issue have not been made clear yet.

### 2.4.2. L1-SLR with component-wise updates (L1-SLR-c)

This algorithm uses the prior distribution Eq.(4) rather than the hierarchical prior distributions Eq.(5). Thus the relevance parameters do not appear in the algorithm derivation. The weight updating rule is obtained by directly differentiating the surrogate function.

The MAP estimated of weight parameters are obtained by

$$\mathbf{w} = \arg\max \ \left\{ \log P(\mathbf{y} \mid \mathbf{w}) - \sqrt{\lambda} \left\| \mathbf{w} \right\|_{l_1} \right\}.$$

Taking the lower bound of the $\log P(\mathbf{y} \mid \mathbf{w})$, we have the following surrogate function,

$$Q(\mathbf{w} \mid \hat{\mathbf{w}}^{(t)}) = \mathbf{w}^t (g(\hat{\mathbf{w}}^{(t)}) - B\hat{\mathbf{w}}^{(t)}) + \frac{1}{2}\mathbf{w}^t B\mathbf{w} - \sqrt{\lambda} \left\| \mathbf{w} \right\|_{l_1}$$

The function $g(\cdot)$ is the gradient of $\log P(\mathbf{y} \mid \mathbf{w})$. The matrix $B$ can be any positive-definite matrix that bounds Hessian of $\log P(\mathbf{y} \mid \mathbf{w})$ everywhere. Here we chose

$$B = -X' \begin{bmatrix} 0.25 & & \\ & \ddots & \\ & & 0.25 \end{bmatrix} X$$

$$\textit{where } X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \ : \ D \times N.$$

Then by directly differentiating $Q(\mathbf{w} \mid \hat{\mathbf{w}}^{(t)})$ with respect to $w_d$ we have the following updating rule,

$$w_d^{(t+1)} = soft\left( w_d^{(t+1)} - \frac{g_k(\hat{\mathbf{w}}^{(t)})}{B_{kk}}; -\frac{\sqrt{\lambda}}{B_{kk}} \right) \qquad (24)$$

$$\textit{where } soft(a;\delta) = sign(a)\max\{0, |a| - \delta\}$$

For more mathematical details, please read the original paper.

## 3. Algorithm Summary

The algorithms of SLR-LAP and SLR-VAR are described here. The other algorithms (RLR-LAP, RLR-VAR, RVM, L1-SLR-LAP) are similar therefore omitted.

### - SLR-LAP

**1. Initialization**

Set $\bar{\alpha}_d = 1 \quad d = 1,\cdots,D$

**2. W-step**

$Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S) \;\; where \;\; S \equiv H(\bar{\mathbf{w}})^{-1}$

$\bar{\mathbf{w}}$ is the maximizer of $E(\mathbf{w})$ that can be obtained by the Newton-Rapson method.

$$E(\mathbf{w}) = \sum_{n=1}^{N} \{ y_n \log \sigma_n + (1-y_n)\log(1-\sigma_n) \} - \frac{1}{2}\mathbf{w}^t \bar{A}\mathbf{w} + const$$

where

$$\begin{cases} \dfrac{\partial E}{\partial \mathbf{w}} = X\boldsymbol{\delta} - \bar{A}\mathbf{w} \\ \dfrac{\partial^2 E}{\partial \mathbf{w} \partial \mathbf{w}^t} = -XBX' - \bar{A} \equiv -H(\mathbf{w}) \end{cases},$$

$\bar{A} = diag(\bar{\alpha}_1, \cdots, \bar{\alpha}_D) \;\; : \; D \times D$

$\boldsymbol{\delta} = [y_1 - \sigma_1, \cdots, y_N - \sigma_N]^t \; : \; N \times 1$

$X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \; : \; D \times N$

$B = diag(\sigma_1(1-\sigma_1), \cdots, \sigma_N(1-\sigma_N)) \; : \; N \times N$

**3. A-step:**

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d)$$

$\bar{\alpha}_d = \dfrac{1 - \bar{\alpha}_d s_d^2}{\bar{w}_d^2}, \; \bar{\gamma}_d = \dfrac{1}{2}$ ( see original rule : $\bar{\alpha}_d = \dfrac{1}{\bar{w}_d^2 + s_d^2}, \; \bar{\gamma}_d = \dfrac{1}{2}$)

$\bar{w}_d, s_d^2$ are the *d*th element and the *d*th diagonal element of $\bar{\mathbf{w}}$ and $S$, respectively.

**4. Convergence:**

Continue W-step and A-step alternately until the change of weight parameters is very small or the number of iterations exceeds the predefined value.

## - SLR-VAR

### 1. Initialization

Set $\bar{\alpha}_d = 1 \quad d = 1,\cdots,D$ and $\xi_n = 2 \quad n = 1,\cdots,N$

### 2. W-step

$$Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S)$$

$$\begin{cases} S^{-1} = \bar{A} + \sum_{n=1}^{N} 2\lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^t = \bar{A} + 2X\Lambda X^t : D \times D \\ \bar{\mathbf{w}} = S\sum_{n=1}^{N}(y_n - 0.5)\mathbf{x}_n \quad : \quad D \times 1 \\ \quad = \begin{cases} (\bar{A} + 2X\Lambda X^t)^{-1}X\tilde{\mathbf{y}} & D < N \\ \bar{A}^{-1}X(\Lambda^{-1} + 2X^t\bar{A}^{-1}X)^{-1}\Lambda^{-1}\tilde{\mathbf{y}} & D > N \end{cases} \end{cases}$$

$$\tilde{\mathbf{y}} = [y_1 - 0.5,\cdots,y_N - 0.5]^t \ : \ N \times 1$$

$$X = [\mathbf{x}_1,\cdots,\mathbf{x}_N] \ : \ D \times N$$

$$\Lambda = diag(\lambda(\xi_1),\cdots,\lambda(\xi_N)) \ : \ N \times N$$

$$\lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right) > 0 \ \ (\xi > 0)$$

### 3. A-step:

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D} \Gamma(\alpha_d; \bar{\alpha}_d, \bar{\gamma}_d)$$

$$\bar{\alpha}_d = \frac{1 - \bar{\alpha}_d s_d^2}{\bar{w}_d^2}, \ \bar{\gamma}_d = \frac{1}{2} \quad \text{( see original rule : } \bar{\alpha}_d = \frac{1}{\bar{w}_d^2 + s_d^2}, \ \bar{\gamma}_d = \frac{1}{2})$$

$\bar{w}_d, s_d^2$ are the $d$th element and the $d$th diagonal element of $\bar{\mathbf{w}}$ and $S$, respectively.

### 4. ξ -step:

$$\xi_n^2 = \mathbf{x}_n^t(\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)\mathbf{x}_n \qquad n = 1,\cdots,N$$

### 5. Convergence:

Continue W-step, A-step and ξ -Step alternately until the change of weight parameters is very small or the number of iterations exceeds the predefined value.

# References

- Jaakkola TS, Jordan MI (2000), Bayesian parameter estimation via variational methods, *Statistics and Computing*, 10, pp.25—37
- Bishop C, Tipping ME (2000),Variational relevance vector machines, *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pp.46-53
- Bishop C (2006), Pattern recognition and machine learning, Springer, New York
- Tipping ME(2001), Sparse Bayesian Learning and the Relevance Vector Machine, *J Machine Learning Research*, 1, pp.211-244
- Krishnapuram B, Hartemink AJ, Carin L and Figueiredo MAT (2004), A Bayesian Approach to Joint Feature Selection and Classifier Design, IEEE Trans. Pattern Analysis and Machine Intelligence, 26, pp.1105-1111
- Krishnapuram B, Carin L, Figueiredo MAT, and Hartemink AJ (2005), Sparse Multinomial Logistic Regression Fast Algorithms and Generalization Bounds, IEEE Trans. Pattern Analysis and Machine Intelligence, 27, pp.957-968
- MacKay D (1992), Bayesian interpolation, Neural Computation, 4, pp.415-447

## Appendix：Deformation of formula：Details

This appendix presents details of equation manipulation.

### In 1.1 Sparse Logistic Regression (SLR)

$$
\begin{aligned}
P_0(w) &= \int P_0(w \mid \alpha) P_0(\alpha) d\alpha \\
&= \int (2\pi)^{-\frac{1}{2}} \alpha^{\frac{1}{2}} \exp\left(-\frac{\alpha}{2} w^2\right) \cdot \alpha^{-1} d\alpha \\
&= (2\pi)^{-\frac{1}{2}} \int \alpha^{-\frac{1}{2}} \exp\left(-\frac{\alpha}{2} w^2\right) d\alpha \\
&= (2\pi)^{-\frac{1}{2}} \Gamma(1/2) \left(\frac{w^2}{2}\right)^{-\frac{1}{2}} \int \frac{1}{\Gamma(1/2)} \left(\frac{w^2}{2}\right)^{\frac{1}{2}} \alpha^{-\frac{1}{2}} \exp\left(-\frac{\alpha}{2} w^2\right) d\alpha \\
&= (2\pi)^{-\frac{1}{2}} \pi^{\frac{1}{2}} \left(\frac{w^2}{2}\right)^{-\frac{1}{2}} \cdot 1 \\
&= \frac{1}{|w|}
\end{aligned}
$$

### In 1.4 L1-Sparse Logistic Regression (L1-SLR)

$$
\begin{aligned}
P_0(w) &= \int P_0(w \mid \alpha) P_0(\alpha) d\alpha \\
&= \int (2\pi\alpha)^{-\frac{1}{2}} \exp\left(-\frac{w^2}{2\alpha}\right) \cdot \frac{\lambda}{2} \exp\left(-\frac{\lambda}{2}\alpha\right) d\alpha \\
&= (2\pi)^{-\frac{1}{2}} \frac{\lambda}{2} \int \alpha^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{w^2}{\alpha} + \lambda\alpha\right)\right\} d\alpha \\
&= \frac{1}{2}\sqrt{\lambda} \exp(-\sqrt{\lambda}\,|w|)
\end{aligned}
$$

Here we used the following formula,

$$
\int x^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{b^2}{x} + ax\right)\right\} dx = \sqrt{\frac{\pi}{a}} \exp(-2b\sqrt{a}) .
$$

### In 2.1.1. SLR-LAP

**W-step:**

$$\begin{aligned}\log Q(\mathbf{w}) &=< \log P(\mathbf{y},\mathbf{w},\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} \\ &= \log P(\mathbf{y}\,|\,\mathbf{w}) + < \log P_0(\mathbf{w}\,|\,\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} + const \\ &= \sum_{n=1}^{N}\{y_n \log\sigma_n + (1-y_n)\log(1-\sigma_n)\} - \frac{1}{2}\mathbf{w}^t\bar{A}\mathbf{w} + const \\ &= E(\bar{\mathbf{w}}) - \frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}})^t H(\bar{\mathbf{w}})(\mathbf{w}-\bar{\mathbf{w}}) + const\end{aligned}$$

$Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S)$ where $S \equiv H(\bar{\mathbf{w}})^{-1}$ and $\bar{\mathbf{w}}$ is maximizer of $E(\mathbf{w})$.

Using the following notations:

$$\bar{A} = diag(\bar{\alpha}_1,\cdots,\bar{\alpha}_D) \quad : D\times D$$

$$E(\mathbf{w}) = \sum_{n=1}^{N}\{y_n \log\sigma_n + (1-y_n)\log(1-\sigma_n)\} - \frac{1}{2}\mathbf{w}^t\bar{A}\mathbf{w} \qquad .$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}} &= \sum_{n=1}^{N}\left\{\frac{y_n}{\sigma_n}\frac{\partial\sigma_n}{\partial\mathbf{w}} - \frac{1-y_n}{1-\sigma_n}\frac{\partial\sigma_n}{\partial\mathbf{w}}\right\} - \bar{A}\mathbf{w} \\ &= \sum_{n=1}^{N}\{y_n(1-\sigma_n) - (1-y_n)\sigma_n\}\mathbf{x}_n - \bar{A}\mathbf{w} \quad , \\ &= \sum_{n=1}^{N}(y_n - \sigma_n)\mathbf{x}_n - \bar{A}\mathbf{w} \\ &= X\boldsymbol{\delta} - \bar{A}\mathbf{w}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 E}{\partial\mathbf{w}\partial\mathbf{w}^t} &= \sum_{n=1}^{N} -\mathbf{x}_n\frac{\partial\sigma_n}{\partial\mathbf{w}^t} - \bar{A} \\ &= -\sum_{n=1}^{N}(1-\sigma_n)\sigma_n\mathbf{x}_n\mathbf{x}_n^t - \bar{A} \\ &= -XBX' - \bar{A} \\ &\equiv -H(\mathbf{w})\end{aligned}$$

$$\begin{aligned}&\boldsymbol{\delta} = [y_1-\sigma_1,\cdots,y_N-\sigma_N]^t \quad : N\times 1 \\ &X = [\mathbf{x}_1,\cdots,\mathbf{x}_N] \quad : D\times N \qquad . \\ &B = diag(\sigma_1(1-\sigma_1),\cdots,\sigma_N(1-\sigma_N)) \quad : N\times N\end{aligned}$$

**A-step:**

$$\begin{aligned}\log Q(\boldsymbol{\alpha}) &=< \log P(\mathbf{y},\mathbf{w},\boldsymbol{\alpha}) >_{Q(\mathbf{w})} \\ &=< \log P_0(\mathbf{w}\,|\,\boldsymbol{\alpha}) >_{Q(\mathbf{w})} + \log P_0(\boldsymbol{\alpha}) + const \\ &= -\frac{1}{2}\sum_{d=1}^{D}\left(\alpha_d < w_d^2 > -\log\alpha_d\right) - \sum_{d=1}^{D}\log\alpha_d + const \\ &= -\frac{1}{2}\sum_{d=1}^{D}\left(\alpha_d < w_d^2 > +\log\alpha_d\right) + const \\ &\equiv \sum_{d=1}^{D}\log Q(\alpha_d)\end{aligned}$$

$$Q(\boldsymbol{\alpha}) = \prod_{d=1}^{D} Q(\alpha_d) = \prod_{d=1}^{D}\Gamma(\alpha_d;\bar{\alpha}_d,\bar{\gamma}_d) \text{ where}$$

$$\frac{\bar{\gamma}_d}{\bar{\alpha}_d} = \frac{\bar{w}_d^2 + s_d^2}{2}, \bar{\gamma}_d = \frac{1}{2} \quad \Rightarrow \bar{\alpha}_d = \frac{1}{\bar{w}_d^2 + s_d^2} \quad \left(\bar{\gamma}_d = \frac{1}{2}\right)$$

$\bar{w}_d, s_d^2$ are the *d*th element and the *d*th diagonal element of $\bar{\mathbf{w}}$ and $S$, respectively.

Fast convergence rule can be obtained at the convergent point.

$$\begin{aligned}&\bar{\alpha}_d = \frac{1}{\bar{w}_d^2 + s_d^2} \\ &\bar{\alpha}_d\bar{w}_d^2 + \bar{\alpha}_d s_d^2 = 1 \\ &\bar{\alpha}_d\bar{w}_d^2 = 1 - \bar{\alpha}_d s_d^2 \\ &\bar{\alpha}_d = \frac{1-\bar{\alpha}_d s_d^2}{\bar{w}_d^2} \quad \left(\bar{\gamma}_d = \frac{1}{2}\right).\end{aligned}$$

<u>In 2.1.2. SLR with variational approximation (SLR-VAR)</u>

From the variational approximation to the logistic function(Jaakkola and Jordan 2000), we have

$$\sigma(x) \geq \sigma(\xi)\exp\left\{\frac{x-\xi}{2} - \lambda(\xi)(x^2 - \xi^2)\right\}$$

where $\lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right) > 0 \ (\xi > 0)$. Since the likelihood function can be rewritten as

$$\begin{aligned}P(\mathbf{y}\,|\,\mathbf{w}) &= \prod \sigma_n^{y_n}(1-\sigma_n)^{1-y_n} \\ &= \prod \sigma_n^{y_n}(\exp(-z_n)\sigma_n)^{1-y_n} \\ &= \prod \sigma_n \exp(-z_n(1-y_n))\end{aligned}$$

then

$$P(\mathbf{y}\mid\mathbf{w}) \ge \prod \sigma(\xi_n)\exp\left\{\frac{z_n-\xi_n}{2} - \lambda(\xi_n)(z_n^2-\xi_n^2) - z_n(1-y_n)\right\} \equiv h(\mathbf{y},\mathbf{w},\boldsymbol{\xi})$$

where $z_n = \mathbf{x}_n^t\mathbf{w}$.

Using this formula, FE is lower bounded as follows

$$FE(Q(\mathbf{w},\boldsymbol{\alpha})) \ge \int Q(\mathbf{w},\boldsymbol{\alpha})\log\frac{h(\mathbf{y},\mathbf{w},\boldsymbol{\xi})P_0(\mathbf{w}\mid\boldsymbol{\alpha})P_0(\boldsymbol{\alpha})}{Q(\mathbf{w},\boldsymbol{\alpha})}d\boldsymbol{\alpha}d\mathbf{w} \equiv FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi}).$$

$FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi})$ is maximized with respect to $Q(\mathbf{w}),Q(\boldsymbol{\alpha}),\xi$ alternately.

<u>W-step:</u>

$$\begin{aligned}
\log Q(\mathbf{w}) &=< \log h(\mathbf{y},\mathbf{w},\boldsymbol{\xi}) + \log P_0(\mathbf{w}\mid\boldsymbol{\alpha}) + \log P_0(\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} \\
&= \log h(\mathbf{y},\mathbf{w},\boldsymbol{\xi}) + < \log P_0(\mathbf{w}\mid\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} + const \\
&= \sum_{n=1}^{N}\left\{\log\sigma(\xi_n) + \frac{z_n-\xi_n}{2} - \lambda(\xi_n)(z_n^2-\xi_n^2) - z_n(1-y_n)\right\} - \frac{1}{2}\mathbf{w}^t\bar{A}\mathbf{w} + const \\
&= \frac{1}{2}\sum_{n=1}^{N}\left\{(2y_n-1)z_n - 2\lambda(\xi_n)z_n^2\right\} - \frac{1}{2}\mathbf{w}^t\bar{A}\mathbf{w} + const \\
&= \frac{1}{2}\sum_{n=1}^{N}\left\{(2y_n-1)\mathbf{x}_n^t\mathbf{w} - 2\lambda(\xi_n)\mathbf{w}^t\mathbf{x}_n\mathbf{x}_n^t\mathbf{w}\right\} - \frac{1}{2}\mathbf{w}^t\bar{A}\mathbf{w} + const \\
&= -\frac{1}{2}\mathbf{w}^t\left(\bar{A} + \sum_{n=1}^{N}2\lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^t\right)\mathbf{w} + \left(\sum_{n=1}^{N}(y_n-0.5)\mathbf{x}_n^t\right)\mathbf{w} + const \\
&= -\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}})^t S^{-1}(\mathbf{w}-\bar{\mathbf{w}}) + const
\end{aligned}$$

Thus $Q(\mathbf{w}) \sim N(\bar{\mathbf{w}},S)$ where

$$\begin{cases}
S^{-1} = \bar{A} + \sum_{n=1}^{N}2\lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^t = \bar{A} + 2X\Lambda X^t \ : D\times D \\
\quad \bar{\mathbf{w}} = S\sum_{n=1}^{N}(y_n-0.5)\mathbf{x}_n \\
\qquad = (\bar{A} + 2X\Lambda X^t)^{-1}X\tilde{\mathbf{y}} \\
\qquad = \bar{A}^{-1}X(\Lambda^{-1} + 2X^t\bar{A}^{-1}X)^{-1}\Lambda^{-1}\tilde{\mathbf{y}} \ : \ D\times 1
\end{cases}$$

$$\begin{aligned}
&\tilde{\mathbf{y}} = [y_1-0.5,\cdots,y_N-0.5]^t \ : \ N\times 1 \\
&X = [\mathbf{x}_1,\cdots,\mathbf{x}_N] \ : \ D\times N \\
&\Lambda = diag(\lambda(\xi_1),\cdots,\lambda(\xi_N)) \ : \ N\times N
\end{aligned}$$

<u>$\boldsymbol{\xi}$ -Step:</u>

$$
\begin{aligned}
FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi}) &=< \log h(\mathbf{y},\mathbf{w},\boldsymbol{\xi}) + \log P_0(\mathbf{w}\,|\,\boldsymbol{\alpha}) + \log P_0(\boldsymbol{\alpha}) >_{Q(\mathbf{w})Q(\boldsymbol{\alpha})} \\
&=< \log h(\mathbf{y},\mathbf{w},\boldsymbol{\xi}) >_{Q(\mathbf{w})Q(\boldsymbol{\alpha})} + const \\
&=< \sum_{n=1}^{N} \left\{ \log \sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)(z_n^2 - \xi_n^2) \right\} >_{Q(\mathbf{w})} + const
\end{aligned}
$$

Taking partial derivative of $FE(Q(\mathbf{w},\boldsymbol{\alpha}),\boldsymbol{\xi})$, we have

$$
\begin{aligned}
\frac{\partial FE}{\partial \xi_n} &= \frac{\sigma'(\xi_n)}{\sigma(\xi_n)} - \frac{1}{2} - \lambda'(\xi_n)(< z_n^2 > - \xi_n^2) + \lambda(\xi_n)\cdot 2\xi_n \\
&= 1 - \sigma(\xi_n) - \frac{1}{2} - \lambda'(\xi_n)(< z_n^2 > - \xi_n^2) + \sigma(\xi_n) - \frac{1}{2} \\
&= -\lambda'(\xi_n)(< z_n^2 > - \xi_n^2)
\end{aligned}
$$

Since $\lambda'(\xi_n) \neq 0$ for $\xi_n > 0$, $\partial FE \big/ \partial \xi_n = 0$ leads to

$$
\begin{aligned}
\xi_n^2 &=< z_n^2 > \\
&= \mathbf{x}_n^t < \mathbf{w}\mathbf{w}^t > \mathbf{x}_n \quad . \\
&= \mathbf{x}_n^t(\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)\mathbf{x}_n
\end{aligned}
$$

By noticing that $\xi_n^2$ is the $n$th diagonal element of $X^t(\bar{\mathbf{w}}\bar{\mathbf{w}}^t + S)X$ and

$$
\begin{aligned}
X^t S X &= X^t(\bar{A} + 2X\Lambda X^t)^{-1}X \\
&= X^t\left(\bar{A}^{-1} - \bar{A}^{-1}X((2\Lambda)^{-1} + X^t\bar{A}^{-1}X)^{-1}X^t\bar{A}^{-1}\right)X \\
&= X^t\bar{A}^{-1}X\left(I - ((2\Lambda)^{-1} + X^t\bar{A}^{-1}X)^{-1}X^t\bar{A}^{-1}X\right) \quad , \\
&= X^t\bar{A}^{-1}X\left((2\Lambda)^{-1} + X^t\bar{A}^{-1}X\right)^{-1}(2\Lambda)^{-1}
\end{aligned}
$$

we have $\xi_n^2 = \left[ X^t\bar{\mathbf{w}}\bar{\mathbf{w}}^t X + X^t\bar{A}^{-1}X\left((2\Lambda)^{-1} + X^t\bar{A}^{-1}X\right)^{-1}(2\Lambda)^{-1} \right]_{nn}$ .

<u>In 2.2.1 RLR with Laplace approximation (RLR-LAP)</u>

<u>A-step:</u>

$$
\begin{aligned}
\log Q(\alpha) &=< \log P(\mathbf{y},\mathbf{w},\alpha) >_{Q(\mathbf{w})} \\
&=< \log P_0(\mathbf{w}\,|\,\alpha) >_{Q(\mathbf{w})} + \log P_0(\alpha) + const \\
&= -\frac{1}{2} < \mathbf{w}^t\mathbf{w} > \alpha + \frac{D}{2}\log\alpha - \log\alpha + const \\
&= -\frac{1}{2} < \mathbf{w}^t\mathbf{w} > \alpha + \left(\frac{D}{2} - 1\right)\log\alpha + const
\end{aligned}
$$

Thus $Q(\alpha) = \Gamma(\alpha\ ;\bar{\alpha},\bar{\gamma})$ where $\bar{\alpha} = \dfrac{D}{\sum_{d=1}^{D}\left(\bar{w}_d^2 + s_d^2\right)}$, $\bar{\gamma} = \dfrac{D}{2}$ (15)

Fast updating rule： $Q(\alpha) = \Gamma(\alpha\ ;\bar{\alpha},\bar{\gamma})$ where $\bar{\alpha} = \dfrac{D - \bar{\alpha}\sum_{d=1}^{D} s_d^2}{\sum_{d=1}^{D}\bar{w}_d^2}$, $\bar{\gamma} = \dfrac{D}{2}$.

In 2.4.1 L1-SLR with Laplace approximation (L1-SLR-LAP)

W-step

$$
\begin{aligned}
\log Q(\mathbf{w}) &=< \log P(\mathbf{y},\mathbf{w},\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} \\
&= \log P(\mathbf{y}\,|\,\mathbf{w}) + < \log P_0(\mathbf{w}\,|\,\boldsymbol{\alpha}) >_{Q(\boldsymbol{\alpha})} + const \\
&= \sum_{n=1}^{N}\left\{y_n \log\sigma_n + (1-y_n)\log(1-\sigma_n)\right\} - \frac{1}{2}\mathbf{w}^t\bar{V}\mathbf{w} + const \\
&= E(\bar{\mathbf{w}}) - \frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}})^t H(\bar{\mathbf{w}})(\mathbf{w}-\bar{\mathbf{w}}) + const
\end{aligned}
$$

$Q(\mathbf{w}) \sim N(\bar{\mathbf{w}}, S)$ where $S \equiv H(\bar{\mathbf{w}})^{-1}$ and $\bar{\mathbf{w}}$ is maximizer of $E(\mathbf{w})$.

Using the following notations:

$$
\bar{V} = diag(<\alpha_1^{-1}>,\cdots,<\alpha_D^{-1}>)\quad : D\times D
$$

$$
E(\mathbf{w}) = \sum_{n=1}^{N}\left\{y_n \log\sigma_n + (1-y_n)\log(1-\sigma_n)\right\} - \frac{1}{2}\mathbf{w}^t\bar{V}\mathbf{w} \qquad .
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}} &= \sum_{n=1}^{N}\left\{\frac{y_n}{\sigma_n}\frac{\partial\sigma_n}{\partial\mathbf{w}} - \frac{1-y_n}{1-\sigma_n}\frac{\partial\sigma_n}{\partial\mathbf{w}}\right\} - \bar{V}\mathbf{w} \\
&= \sum_{n=1}^{N}\left\{y_n(1-\sigma_n) - (1-y_n)\sigma_n\right\}\mathbf{x}_n - \bar{V}\mathbf{w}\ , \\
&= \sum_{n=1}^{N}(y_n - \sigma_n)\mathbf{x}_n - \bar{V}\mathbf{w} \\
&= X\boldsymbol{\delta} - \bar{V}\mathbf{w}
\end{aligned}
$$

$$\begin{aligned}\frac{\partial^2 E}{\partial \mathbf{w} \partial \mathbf{w}^t} &= \sum_{n=1}^{N} -\mathbf{x}_n \frac{\partial \sigma_n}{\partial \mathbf{w}^t} - \overline{V} \\ &= -\sum_{n=1}^{N} (1-\sigma_n)\sigma_n \mathbf{x}_n \mathbf{x}_n^t - \overline{V} \\ &= -XBX' - \overline{V} \\ &\equiv -H(\mathbf{w})\end{aligned}$$

$$\begin{aligned}&\boldsymbol{\delta} = [y_1 - \sigma_1, \cdots, y_N - \sigma_N]^t \ : N \times 1 \\ &X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \ : D \times N \qquad . \\ &B = diag(\sigma_1(1-\sigma_1), \cdots, \sigma_N(1-\sigma_N)) \ : N \times N\end{aligned}$$