Logistic Regression

- Predict a binary output \( y_t \in \{0, 1\} \) from an input \( x_t \).
- The logistic regression model has the form
  \[ y_t = \sigma(w^\top x_t) + \epsilon_t, \]
  where
  \[ \sigma(\xi) = \frac{1}{1 + e^{-\xi}} = \frac{e^\xi}{1 + e^\xi}. \]
- Model input-output by a conditional Bernoulli distribution
  \[ P(y_t = 1|x_t) = \sigma(w^\top x_t). \]
Logistic Regression: MLE

- Given \( \{(x_t, y_t)\}_{t=1}^{N} \), the likelihood is given by

\[
p(y|X, w) = \prod_{t=1}^{N} p(y_t = 1|x_t)^{y_t} (1 - p(y_t = 1|x_t))^{1-y_t} = \prod_{t=1}^{N} \sigma(w^\top x_t)^{y_t} (1 - \sigma(w^\top x_t))^{1-y_t}.
\]

- Then log-likelihood function is given by

\[
\mathcal{L} = \sum_{t=1}^{N} \log p(y_t|x_t) = \sum_{t=1}^{N} \{y_t \log \sigma_t + (1 - y_t) \log (1 - \sigma_t)\},
\]

where \( \sigma_t = \sigma(w^\top x_t) \).

- This is a nonlinear function of \( w \) whose maximum cannot be computed in a closed form.

- Iterative re-weighted least squares (IRLS) is a popular algorithm, derived from Newton’s method.

Multiclass Logistic Regression

- Model input-output by a softmax transformation of linear functions of inputs:

\[
p(y_t = k|x_t) = \frac{\exp\{w_k^\top x_t\}}{\sum_{k=1}^{K} \exp\{w_j^\top x_t\}}.
\]

- Given \( Y \in \mathbb{R}^{K \times N} \) (each column \( y_t \in \mathbb{R}^K \) follows the 1-of-\( K \) coding) and \( X \in \mathbb{R}^{D \times N} \), the likelihood is given by

\[
p(Y|X, w_1, \ldots, w_K) = \prod_{t=1}^{N} \prod_{k=1}^{K} p(y_t = k|x_t)^{Y_{tk}}.
\]

- The log-likelihood is given by

\[
\mathcal{L} = \sum_{t=1}^{N} \sum_{k=1}^{K} Y_{tk} \log [p(y_t = k|x_t)].
\]

- One can apply Newton’s update to derive IRLS, as in logistic regression.

Logistic Regression: IRLS

- Newton’s update has the form

\[
\Delta w = \left[ \sum_{t=1}^{N} \sigma_t (1 - \sigma_t) x_t x_t^\top \right]^{-1} \sum_{t=1}^{N} (y_t - \sigma_t) x_t,
\]

where

\[
\begin{bmatrix}
\sigma_1(1 - \sigma_1) & 0 \\
0 & \ddots & 0 \\
0 & \ddots & 0 \\
\frac{\sigma_1 - \sigma_N}{\sigma_1(1 - \sigma_1)} & 0 & \cdots & \frac{\sigma_K - \sigma_N}{\sigma_N(1 - \sigma_K)}
\end{bmatrix}
\]

which can be re-written as

\[
\Delta w = \left( X^\top V X \right)^{-1} X^\top V Y^*,
\]

where

\[
V_t = \begin{bmatrix}
\sigma_1(1 - \sigma_1) & 0 \\
0 & \ddots & 0 \\
0 & \ddots & 0 \\
\frac{\sigma_1 - \sigma_N}{\sigma_1(1 - \sigma_1)} & 0 & \cdots & \frac{\sigma_K - \sigma_N}{\sigma_N(1 - \sigma_K)}
\end{bmatrix},
\]

\[
Y^* = \begin{bmatrix}
\frac{\sigma_1 - \sigma_N}{\sigma_1(1 - \sigma_1)} \\
\frac{\sigma_2 - \sigma_N}{\sigma_2(1 - \sigma_2)} \\
\vdots \\
0
\end{bmatrix}.
\]

Laplace Approximation

Consider a distribution over continuous variables

\[
p(x) = \frac{1}{Z} f(x).
\]

Laplace approximation finds a Gaussian approximation \( q(x) \) to \( p(x) \), where \( q(x) \) is centered on a mode of the distribution \( p(x) \):

\[
q(x) = \frac{|A|^\frac{1}{2}}{(2\pi)^{\frac{D}{2}}} \exp\left\{ -\frac{1}{2} (x - \bar{x})^\top A (x - \bar{x}) \right\},
\]

where \( \bar{x} \) is a mode of \( p(x) \), i.e., a point \( \bar{x} \) which satisfies

\[
\frac{df(x)}{dx} \bigg|_{x=\bar{x}} = 0,
\]

and

\[
A = -\nabla^2 \log f(\bar{x}) = -\frac{d}{dx} \left[ \nabla \log f(x) \right]^\top \bigg|_{x=\bar{x}}.
\]
Bayesian Logistic Regression: Laplace Approximation

Consider the 2nd-order Taylor series expansion of \( \log f(x) \) around the mode \( \bar{x} \),

\[
\log f(x) \approx \log f(\bar{x}) + [\nabla \log f(\bar{x})]^{\top}(x - \bar{x}) - \frac{1}{2}(x - \bar{x})^{\top}[-\nabla^{2}\log f(\bar{x})](x - \bar{x})
\]

leading to

\[
f(x) \approx f(\bar{x}) \exp \left\{ -\frac{1}{2}(x - \bar{x})^{\top}A(x - \bar{x}) \right\}.
\]

Then the normalized distribution \( q(x) \) is computed as

\[
q(x) = \frac{|A|^{1/2}}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2}(x - \bar{x})^{\top}A(x - \bar{x}) \right\}.
\]

Bayesian Logistic Regression: Laplace Approximation

- Recall \( y_{t} = \sigma(w^{\top}x_{t}) + \epsilon_{t} \).
- Place a Gaussian prior on \( w \), i.e.,
  \[
p(w) = \mathcal{N}(w | \mu_{0}, \Sigma_{0}).
\]
  where \( \mu_{0} \) and \( \Sigma_{0} \) are hyperparameters.
- Find Laplace approximation to the posterior distribution over \( w \),
  \[
  q(w) \approx p(w|y, X) \propto p(y|X, w)p(w).
  \]
- Compute predictive distribution, given a new input \( x_{*} \),
  \[
p(y_{*} = 1|x_{*}, X) = \int p(y_{*}|x_{*}, w)p(w|y, X)dw
  \approx \int \sigma(w^{\top}x_{*})q(w)dw
  \]

Laplace Approximation: An Example

- The distribution
  \[
p(x) = \frac{1}{Z} \exp\{-x^{2}/2\} \sigma(20x + 4),
  \]
  is shown in yellow.
- Laplace approximation \( q(x) \) centered on the mode of \( p(x) \) is shown in red.

Posterior Distribution

The posterior distribution over \( w \) is given by

\[
p(w|y, X) \propto p(y|X, w)p(w),
\]

where

\[
p(y|X, w) = \prod_{t=1}^{N} \sigma(w^{\top}x_{t})^{y_{t}}(1 - \sigma(w^{\top}x_{t}))^{(1-y_{t})},
\]

\[
p(w) = \mathcal{N}(w | \mu_{0}, \Sigma_{0}).
\]

Thus, the log of posterior distribution is given by

\[
\log p(w|y, X) = -\frac{1}{2}(w - \mu_{0})^{\top}\Sigma_{0}^{-1}(w - \mu_{0})
+ \sum_{t=1}^{N}\{y_{t}\log \sigma(w^{\top}x_{t}) + (1 - y_{t})\log (1 - \sigma(w^{\top}x_{t}))\} + \text{const}.
\]
Gaussian Approximation to Posterior Distribution

A Gaussian approximation has the form

$$q(w) = \mathcal{N}(w \mid w_{MAP}, \Sigma_N),$$

Solve $\nabla_w \log p(w \mid y, X) = 0$ for $w$ to determine $w_{MAP}$ (which is the mean of the Gaussian), where

$$\nabla_w \log p(w \mid y, X) = -\Sigma_N^{-1}(w - \mu_0) + \sum_{t=1}^N \{y_t(1 - \sigma(w^T x_t)) - (1 - y_t)\sigma(w^T x_t)\}x_t.$$ 

The inverse of the negative Hessian of the log-likelihood is given by

$$\Sigma_N^{-1} = -\nabla_w^2 \log p(w \mid y, X) = \Sigma_0^{-1} + \sum_{t=1}^N \sigma(w^T x_t)(1 - \sigma(w^T x_t))x_t x_t^T.$$ 

Predictive Distribution

Predictive distribution, given a new input $x_*$, is computed by

$$p(y_* = 1 \mid y, X, x_*) = \int p(y_* = 1 \mid x_*, w)p(w \mid y, X)dw$$

$$\approx \int \sigma(w^T x_*)q(w)dw.$$ 

Denote $z_* = w^T x_*$, then we have

$$\sigma(w^T x_*) = \int \delta(z_* - w^T x_*)\sigma(z_*)dz_*,$$

leading to

$$\int \sigma(w^T x_*)q(w)dw = \int \sigma(z_*)r(z_*)dz_*,$$

where

$$r(z_*) = \int \delta(z_* - w^T x_*)q(w)dw.$$ 

Predictive Distribution (Cont’d)

Approximate $\sigma(z_*)$ by the probit function $\Phi(\lambda z_*)$ where a suitable value of $\lambda$ is computed, requiring that the two functions have the same slope at the origin, leading to $\lambda = \pi/8$, and $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(z \mid 0, 1)dz$.

Then we approximately compute the predictive distribution by

$$p(y_* = 1 \mid y, X, x_*) = \int p(y_* = 1 \mid x_*, w)p(w \mid y, X)dw$$

$$\approx \int \sigma(w^T x_*)q(w)dw$$

$$= \int \sigma(z_*)\mathcal{N}(z_* \mid \mu_{z_*}, \sigma_{z_*}^2)dz_*$$

$$\approx \int \Phi(\lambda z_*)\mathcal{N}(z_* \mid \mu_{z_*}, \sigma_{z_*}^2)dz_*$$

$$= \Phi \left( \frac{\mu_{z_*}}{\sqrt{\lambda^2 + \sigma_{z_*}^2}} \right)$$

$$\approx \sigma \left( 1 + \frac{\pi}{8} \sigma_{z_*}^2 \right)^{-1} \mu_{z_*}.$$ 

Predictive Distribution (Cont’d)

$$\mu_{z_*} = \int z_* r(z_*)dz_*$$

$$= \int w^T x_* q(w)dw$$

$$= w_{MAP}^T x_*,$$

$$\sigma_{z_*}^2 = \int r(z_*) (z_*^2 - E^2 z_*) dz_*$$

$$= \int q(w) ((w^T x_*)^2 - (w_{MAP}^T x_*)^2) dw$$

$$= x_*^T \Sigma_N x_*.$$
Variational Bayesian Logistic Regression

- A lower bound on the sigmoid logistic function
- Variational posterior distribution over \( w \)
- Optimize the variational bound on the marginal likelihood to estimate variational parameters
- Inference of hyperparameters
- Variational predictive distribution

A Lower Bound on the Sigmoid Logistic

A lower bound on the sigmoid logistic \( \sigma(x) \), which has the functional form of Gaussian, is calculated as

\[
\sigma(x) \geq \sigma(\xi) \exp \left\{ \frac{1}{2} (x - \xi) - \lambda(\xi)(x^2 - \xi^2) \right\},
\]

where

\[
\lambda(\xi) = \frac{1}{4\xi} \tanh \left( \frac{\xi}{2} \right)
\]

\[
= \frac{1}{2\xi} \left\{ \sigma(\xi) - \frac{1}{2} \right\}.
\]

This lower bound is derived using the convex inequality, which is explained in next slide.

Variational Lower Bound on the Logistic

We first consider the log of sigmoid logistic,

\[
\log \sigma(x) = \log \left( \frac{1}{1 + e^{-x}} \right)
\]

\[
= - \log (1 + e^{-x})
\]

\[
= - \log \left( e^{-x/2} \left( e^{x/2} + e^{-x/2} \right) \right)
\]

\[
= \frac{x}{2} - \log \left( e^{x/2} + e^{-x/2} \right)
\]

Note that \( f(x) = -\log \left( e^{x/2} + e^{-x/2} \right) \) is a convex function of \( x^2 \).

Thus, \( f(x) \) satisfies the following inequality:

\[
f(x) \geq f(\xi) + \frac{\partial f(x)}{\partial x^2} \bigg|_{x^2 = \xi^2} (x^2 - \xi^2),
\]

where \( \frac{\partial f(x)}{\partial x^2} \) is calculated as

\[
\frac{\partial f(x)}{\partial x^2} = \frac{\partial f(x)}{\partial z} \bigg|_{z = x^2}
\]

\[
= \frac{\partial f(x)}{\partial x} \frac{\partial x}{\partial x} \frac{\partial x}{\partial z}
\]

\[
= - \frac{1}{4\xi} \tanh \left( \frac{x}{2} \right).
\]

Using this result, we have

\[
f(x) \geq -\frac{\xi}{2} + \log \sigma(\xi) - \frac{1}{4\xi} \tanh \left( \frac{\xi}{2} \right)(x^2 - \xi^2).
\]
Define $\lambda(\xi) = \frac{1}{2} \tanh \left( \frac{\xi}{2} \right)$. Then we have

$$f(x) \geq -\frac{\xi}{2} + \log \sigma(\xi) - \lambda(\xi)(x^2 - \xi^2),$$

$$\log \sigma(x) - \frac{x}{2} \geq -\frac{\xi}{2} + \log \sigma(\xi) - \lambda(\xi)(x^2 - \xi^2),$$

$$\log \sigma(x) \geq \frac{1}{2}(x - \xi) + \log \sigma(\xi) - \lambda(\xi)(x^2 - \xi^2),$$

leading to the lower bound on the sigmoid logistic

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

To this end, we first consider the single factor the likelihood

$$p(y_t|X, w) = \sigma(z_t)^{y_t} (1 - \sigma(z_t))^{1-y_t} \quad (z_t = w^\top x_t)$$

$$= \left( \frac{1}{1 + e^{-z_t}} \right)^{y_t} \left( 1 - \frac{1}{1 + e^{-z_t}} \right)^{1-y_t}$$

$$= \left( \frac{1}{1 + e^{-z_t}} \right)^{y_t} \left( 1 - \frac{1}{1 + e^{-z_t}} \right)^{1-y_t}$$

$$= \left( \frac{1 + e^{-z_t}}{1 + e^{-z_t}} \right)^{y_t} \left( \frac{e^{-z_t}}{1 + e^{-z_t}} \right)^{1-y_t}$$

$$= e^{z_t y_t} \sigma(-z_t).$$

We make use of a variational approximation based on the local bounds on the sigmoid logistic to compute a variational posterior distribution $q(w) \leq p(w|y, X)$.

- Posterior distribution is computed as

$$p(w|y, X) \propto p(y|X, w) p(w).$$

- Gaussian bounds on the sigmoid logistic

$$\log \sigma(x) - \frac{x}{2} \geq -\frac{\xi}{2} + \log \sigma(\xi) - \lambda(\xi)(x^2 - \xi^2),$$

- Gaussian

Variational posterior distribution $q(w)$ is calculated as

$$p(y|X, w) p(w) \geq h(w, \xi) \propto N(w | \mu_N, \Sigma_N).$$

We make use of the variational lower bound on the sigmoid logistic

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

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

Then we have

$$p(y_t|X, w) = e^{z_t y_t} \sigma(z_t)$$

$$\geq e^{z_t y_t} \sigma(\xi_t) \exp \left\{ -\frac{(z_t + \xi_t)}{2} - \lambda(\xi_t)(z_t^2 - \xi_t^2) \right\}$$

$$= h(w, \xi_t).$$

$$h(w, \xi_t) = e^{z_t y_t} \sigma(\xi_t) \exp \left\{ -\frac{(z_t + \xi_t)}{2} - \lambda(\xi_t)(z_t^2 - \xi_t^2) \right\},$$

where $z_t = w^\top x_t$. 
We have the following lower bound on the joint distribution of $y$ and $w$, conditioned on $X$:

$$p(y, w|X) = p(y|w, X)p(w)$$

$$\geq h(w, \xi)p(w)$$

$$= q(w).$$

$$h(w, \xi) = \prod_{t=1}^{N} e^{z_y^t} \left( - \frac{(z_t + \xi_t)}{2} - \lambda(\xi_t)(z_t^2 - \xi_t^2) \right)$$

$$= \prod_{t=1}^{N} \sigma(\xi_t) \exp \left\{ \frac{-z_t^2 + \xi_t^2}{2} - \lambda(\xi_t) \left( (w^T x_t)^2 - \xi_t^2 \right) \right\}.$$

**Variational Posterior vs. Laplace Approximation**

- **Variational posterior** is computed as:
  $$q(w) = \mathcal{N}(w | \mu_N, \Sigma_N),$$
  $$\mu_N = \Sigma_N^{-1} \Sigma_0^{-1} \mu_0 + \sum_{t=1}^{N} \left( y_t - \frac{1}{2} \right) x_t,$$
  $$\Sigma_N^{-1} = \Sigma_0^{-1} + 2 \sum_{t=1}^{N} \lambda(\xi_t)x_t x_t^T.$$

- **Laplace approximation** yields:
  $$\tilde{q}(w) = \mathcal{N}(w | \hat{\mu}_N, \tilde{\Sigma}_N),$$
  $$\hat{\mu}_N = w_{MAP},$$
  $$\tilde{\Sigma}_N^{-1} = \Sigma_0^{-1} + \sum_{t=1}^{N} \sigma(w^T x_t) (1 - \sigma(w^T x_t)) x_t x_t^T.$$

- The additional flexibility provided by the variational parameters $\xi_t$ leads to the improved accuracy in the approximation.

$$\log [h(w, \xi)p(w)] = \log p(w) + \sum_{t=1}^{N} \left[ \log \sigma(\xi_t) + w^T x_t y_t - (w^T x_t + \xi_t)^2 \right]$$

$$- \lambda(\xi_t) \left( (w^T x_t)^2 - \xi_t^2 \right)$$

$$= - \frac{1}{2} (w - \mu_0)^T \Sigma_0 (w - \mu_0)$$

$$+ \sum_{t=1}^{N} \left[ w^T x_t \left( y_t - \frac{1}{2} \right) - \lambda(\xi_t) w^T x_t x_t^T w \right] + \text{const,}$$

which gives

$$q(w) = \mathcal{N}(w | \mu_N, \Sigma_N),$$

$$\mu_N = \Sigma_N \left( \Sigma_0^{-1} \mu_0 + \sum_{t=1}^{N} \left( y_t - \frac{1}{2} \right) x_t \right),$$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + 2 \sum_{t=1}^{N} \lambda(\xi_t)x_t x_t^T.$$

**Optimizing Variational Parameters**

- We determine the variational parameters by maximizing the lower bound on the marginal likelihood:

$$\log p(y|X) = \log \int p(y|X, w)p(w)dw$$

$$\geq \log \int h(w, \xi)p(w)dw$$

$$= \mathcal{L}(\xi).$$

- We treat $w$ as latent variables and invoke EM:
  - **E-step**: Compute the expected complete-data log-likelihood
    $$Q(\xi, \xi^{old}) = E_{q(w)} \{ \log h(w, \xi)p(w) \},$$
    where variational posterior distribution $q(w)$ is evaluated using $\xi^{old}$.  
  - **M-step**: Re-estimate $\xi$
    $$\xi = \arg \max_{\xi} Q(\xi, \xi^{old}).$$
Algorithm Outline: EM for Variational Parameters

- **E-step:** Compute the variational posterior using $\xi^{\text{old}}$
  $$q(w) = \mathcal{N}(\mathbf{w} | \mu_N, \Sigma_N),$$
  $$\mu_N = \Sigma_N \left( \Sigma_0^{-1} \mu_0 + \sum_{t=1}^N \left( y_t - \frac{1}{2} \mathbf{x}_t \right) \right),$$
  $$\Sigma_N^{-1} = \Sigma_0^{-1} + 2 \sum_{t=1}^N \lambda(\xi_t^{\text{old}}) \mathbf{x}_t \mathbf{x}_t^\top,$$
  $$\lambda(\xi_t^{\text{old}}) = \frac{1}{2 \xi_t^{\text{old}}} \left( \sigma(\xi_t^{\text{old}}) - \frac{1}{2} \right).$$

- **M-step:** Re-estimate $\xi_t$
  $$(\xi_t^{\text{new}})^2 = \mathbf{x}_t^\top (\Sigma_N + \mu_N \mu_N^\top) \mathbf{x}_t.$$

A Lower Bound on the Log Marginal Likelihood

Consider a isotropic Gaussian prior for $\mathbf{w}$, $p(\mathbf{w} | \alpha) = \mathcal{N}(\mathbf{w} | 0, \alpha^{-1} I)$.
We consider a conjugate hyperprior over $\alpha$ by a gamma distribution
$$p(\alpha) = \text{Gam}(\alpha | a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha}.$$

We obtain a lower bound on the log marginal likelihood
$$\log p(y | \mathbf{X}) = \log \int \int p(y | \mathbf{w}, \mathbf{X}) p(\mathbf{w} | \alpha) p(\alpha) \, d\mathbf{w} \, d\alpha \geq \int \int q(\mathbf{w}, \alpha) \log \frac{p(y | \mathbf{w}, \mathbf{X}) p(\mathbf{w} | \alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \, d\mathbf{w} \, d\alpha = \mathcal{F}(q) \quad (\text{Jensen’s inequality}) \geq \int \int q(\mathbf{w}, \alpha) \log \frac{h(\mathbf{w}, \xi) p(\mathbf{w} | \alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \, d\mathbf{w} \, d\alpha = \mathcal{L}(q, \xi) \quad (\text{variational lower bound}).$$

Inference of Hyperparameters

- Now we assume that the hyperparameter $\alpha$ in the prior distribution is also unknown, allowing the value of this parameter to be inferred from the data set.
- We infer the hyperparameter $\alpha$ by combining the global and variational approximations into a single framework so as to maintain a lower bound on the marginal likelihood at each stage.
- Recall $p(y, \mathbf{w} | \mathbf{X}) = p(y | \mathbf{w}, \mathbf{X}) p(\mathbf{w}) \geq h(\mathbf{w}, \xi) p(\mathbf{w}).$
- We consider a variational lower bound $\mathcal{L}(q, \xi)$ on a lower bound $\mathcal{F}(q)$ on the log marginal likelihood $\log p(y | \mathbf{X})$
  $$\log p(y | \mathbf{X}) \geq \mathcal{F}(q) \geq \mathcal{L}(q, \xi) = \int \int q(\mathbf{w}, \alpha) \log \left[ \frac{h(\mathbf{w}, \xi) p(\mathbf{w} | \alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \right] \, d\mathbf{w} \, d\alpha,$$
where we assume $q(\mathbf{w}, \alpha) = q(\mathbf{w}) q(\alpha)$.

Estimation of $q(\mathbf{w}), q(\alpha), \xi$

- We assume that the variational distribution $q(\mathbf{w}, \alpha)$ factorizes between parameters and hyperparameters
  $$q(\mathbf{w}, \alpha) = q(\mathbf{w}) q(\alpha).$$
- Compute $q(\mathbf{w})$ and $q(\alpha)$ using $\xi^{\text{old}}$
  $$\log q(\mathbf{w}) = \mathbb{E}_\alpha \{ \log h(\mathbf{w}, \xi) p(\mathbf{w} | \alpha) p(\alpha) \} + \text{const},$$
  $$\log q(\alpha) = \mathbb{E}_\mathbf{w} \{ \log h(\mathbf{w}, \xi) p(\mathbf{w} | \alpha) p(\alpha) \} + \text{const}.$$
- Determine $\xi$ by maximizing $\mathcal{L}(q, \xi)$. 
Algorithm Outline

- Compute \( q(w) \) and \( q(\alpha) \) using \( \xi^{old} \)

\[
q(w) = \mathcal{N}(w | \mu_N, \Sigma_N), \quad q(\alpha) = \text{Gam}(\alpha | a_N, b_N),
\]

\[
\Sigma_N^{-1}\mu_N = \sum_{t=1}^{N} \left( y_t - \frac{1}{2} \right) x_t, \quad \Sigma_N^{-1} = \mathbb{E}\{\alpha}\mathbb{I} + 2 \sum_{t=1}^{N} \lambda(\xi_t)x_t x_t^T,
\]

\[
a_N = a_0 + \frac{D}{2}, \quad b_N = b_0 + \frac{1}{2} \mathbb{E}\{w^T w\},
\]

\[
\mathbb{E}\{\alpha\} = \frac{a_N}{b_N}, \quad \mathbb{E}\{w^T w\} = \Sigma_N + \mu_N^T\mu_N.
\]

- Compute new \( \xi \)

\[
(\xi_t^{new})^2 = x_t^T (\Sigma_N + \mu_N^T\mu_N) x_t.
\]