S1: Derivations of the Update Equations
Supplementary Material for
A Bayesian Model for Exploiting Application
Constraints to Enable Unsupervised Training of
a P300 based BCI

1 The proposed model

In the paper we have proposed the following model:

\[
\begin{align*}
p(w) &= \mathcal{N}(\mu_w, \Sigma_w) \\
p(z_n^\omega = k) &= \frac{1}{K}, \\
p(t_n^{\omega, r, k} = 1|z_n^\omega) &= \begin{cases} 
1 & z_n^\omega = k \\
0 & z_n^\omega \neq k
\end{cases}, \\
p(t_n^{\omega, r, k} = -1|z_n^\omega) &= \begin{cases} 
0 & z_n^\omega = k \\
1 & z_n^\omega \neq k
\end{cases}, \\
p(x|t_n^{\omega, r, k}w, \beta) &= \mathcal{N}(x_n^{\omega, r, k}w|t_n^{\omega, r, k}, \beta^{-1}).
\end{align*}
\]

In order to simplify the equations, we will drop the superscript \(\omega\) and use 2N pseudo-labels instead.

2 Derivations of update Equations for EM

We want to find update equations for \(w, \beta\) (and \(\alpha\)) that maximizes:

\[
\sum_z p(z|X, w_p, \beta_p) \log p(X, z|w, \beta) + \log p(w|\mu_w, \Sigma_w),
\]

where the full model from above is used. To optimize for a specific parameter, we take the derivative with respect to that parameter, set the resulting equation
equal to zero and solve for that parameter. But to keep the discussion of the update equations readable, we have to simplify:

\[ \sum_{z} p(z|X, w_p, \beta_p) \log p(X, z|\theta) \]  

(2)

into something which is simpler to use in the derivation. We will start by using the following notation: \( \theta_p = \{\beta_p, w_p\} \) and \( \theta = \{\beta, w\} \) which reduces Equation 2 to:

\[ \sum_{z} p(z|X, \theta_p) \log p(X, z|\theta) \]  

(3)

We continue by expanding Equation: 3 by making use of the fact that \( p(z|X, \theta_p) \) and \( p(X, z|\theta) \) are both a product over all rows and columns giving us a product over \( 2N \) different indicator variables:

\[
\sum_{z} p(z|X, \theta_p) \log p(X, z|\theta) \\
= \sum_{z_1} \sum_{z_2} \ldots \sum_{z_{2N}} \prod_{n=1}^{2N} p(z_n|X_n, \theta_p) \sum_{m=1}^{2N} \log p(X_m, z_m|\theta) \\
= \sum_{z_1} p(z_1|X_1, \theta_p) \sum_{z_2} p(z_2|X_2, \theta_p) \ldots \sum_{z_{2N}} p(z_{2N}|X_{2N}, \theta_p) \sum_{m=1}^{2N} \log p(X_m, z_m|\theta) \\
= \sum_{z_1} p(z_1|X_1, \theta_p) \log p(X_1, z_1|\theta) \\
+ \sum_{z_1} p(z_1|X_1, \theta_p) \sum_{z_2} p(z_2|X_2, \theta_p) \ldots \sum_{z_{2N}} p(z_{2N}|X_{2N}, \theta_p) \sum_{m=2}^{2N} \log p(X_m, z_m|\theta) \\
= \sum_{n=1}^{2N} \sum_{z_n} p(z_n|X_n, \theta_p) \log p(X_n, z_n|\theta).
\]

This sum of sums makes the derivations of Equation: 2 with respect to \( w \) and \( \beta \) much less complicated because only:

\[ p(X_n, z_n|\theta) = p(X_n, z_n|w, \beta) \]  

(4)

depends on \( w \) and \( \beta \).

Analyzing Equation: 4 allows us to further simplify the optimization process:

\[
p(X_n, z_n|w, \beta) = p(z_n) \sum_{T_n} p(T_n|z_n)p(X_n|T_n, w, \beta) = p(z_n)p(X_n|\tilde{T}_n(z_n), w, \beta),
\]

2
where $T_n$ are the labels assigned to the individual EEG sequences. Due to the distribution on $t_{n,e,k} | z_{n,e,k}$ there exists one unique assignment for $T_n$ which has probability mass 1. We will indicate this assignment by $\tilde{T}_n(z_n)$. All the other assignments have zero probability mass. This result can be plugged into Equation: 2:

$$
\sum_z p(z|X, w_p, \beta_p) \log p(X, z|w, \beta) = \\
\sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \log p(X_n|\tilde{T}_n(z_n), w, \beta) \\
+ \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \log p(z_n),
$$

from we see that only the first term in the sum depends on $w$ and $\beta$. This means that when we take the derivative of Equation: 2 with respect to $\beta$ or $w$ the second part vanishes and only the derivative of:

$$
\sum_z p(z|X, w_p, \beta_p) \log p(X, z|w, \beta)
$$

has to be found.

2.1 The update equation for $w$:

The previous section allows us to quickly find the update equation for $w$. The derivative of Equation: 1 with respect to $w$ is:

$$
\frac{d}{dw} \sum_z p(z|X, w_p, \beta_p) \log p(X, z|w, \beta) + \frac{d}{dw} \log p(w|\mu_w, \Sigma_w)
$$

$$
= \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \frac{d}{dw} \log p(X_n|\tilde{T}_n(z_n), w, \beta) + \frac{d}{dw} \log p(w|\mu_w, \Sigma_w),
$$

where we have used the result from Equation: 5 to simplify the first term.

The derivation of $\log p(X_n|\tilde{T}_n, w, \beta)$ with respect to $w$ can be found as follows:

3
\[
\frac{d}{dw} \log p(X_n|z_n, w, \beta) \\
= \sum_{r=1}^{R_n} \sum_{k=1}^{K} \frac{d}{dw} \log N(x_{n,r,k} w|\tilde{t}_{n,r,k}(z_n), \beta^{-1}) \\
= \sum_{r=1}^{R_n} \sum_{k=1}^{K} \frac{d}{dw} \left[ -\frac{1}{2} \beta (x_{n,r,k} w - \tilde{t}_{n,r,k}(z_n))^2 \right] \\
= \sum_{r=1}^{R_n} \sum_{k=1}^{K} \left[ -\beta (x_{n,r,k} w - \tilde{t}_{n,r,k}(z_n)) x_{n,r,k}^T \right] \\
= -\beta \left( X_n^T X_n w - X_n^T \tilde{T}_n(z_n) \right).
\]

The value for \( \frac{d}{dw} \log p(w|\mu_w, \Sigma_w) \) is the following:

\[
\frac{d}{dw} \log N(\mu_w, \Sigma_w) \\
= \frac{d}{dw} -\frac{1}{2} (w - \mu_w)^T \Sigma_w^{-1} (w - \mu_w) \\
= -\Sigma_w^{-1} (w - \mu_w).
\]

We can use these results and plug them into the equation above:

\[
= -2N \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \beta \left( X_n^T X_n w - X_n^T \tilde{T}_n(z_n) \right) - \Sigma_w^{-1} (w - \mu_w) \\
= 2N \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \beta \left( X_n^T X_n w - X_n^T \tilde{T}_n(z_n) \right) + \Sigma_w^{-1} w - \Sigma_w^{-1} \mu_w \\
= 2N \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \beta \left( X_n^T + \frac{\Sigma_w^{-1}}{\beta} \right) w \\
= -2N \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \beta \left( X_n^T \tilde{T}_n(z_n) \right) - \Sigma_w^{-1} \mu_w \\
\beta \left( X^T + \frac{\Sigma_w^{-1}}{\beta} \right) w = \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \beta \left( X_n^T \tilde{T}_n(z_n) \right) + \Sigma_w^{-1} \mu_w \\
\left( X^T + \frac{\Sigma_w^{-1}}{\beta} \right) w = \sum_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \left( X_n^T \tilde{T}_n(z_n) \right) + \frac{\Sigma_w^{-1}}{\beta} \mu_w
\]

Where we obtain the updated value \( w \) by setting the above equal to zero and solving for \( w \):
\[ w = \left( \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) X_n^T X_n + \frac{\Sigma_w}{\beta} \right)^{-1} \]
\[ \left( \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) X_n^T \tilde{T}_n(z_n) + \frac{\Sigma_w}{\beta} \mu_w \right), \]

by using:
\[ \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) = 1, \]
we can simplify the update for \( w \):

\[ w = \left( X^T X + \frac{\Sigma_w}{\beta} \right)^{-1} \left( \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) X_n^T \tilde{T}_n(z_n) + \frac{\Sigma_w}{\beta} \mu_w \right). \]

Which shows us that we only need \( 2NK \) different evaluations of some \( z \). This in contrast to \( K^{2N} \) different values for \( z \) in Equation: 2. However rewriting the above as a sum over \( z \) shows us that \( w \) is a linear combination of linear classifiers:

\[ w = \sum_{z} p(z|X, w_p, \beta_p) \left( X^T X + \frac{\Sigma_w}{\beta} \right)^{-1} \left( X^T \tilde{T}(z) + \frac{\Sigma_w}{\beta} \mu_w \right), \]

where the combination weights are the conditional probabilities that the specific \( z \) is correct given the previous parameters and the labels used for training are the ones induced by \( z \).

### 2.2 Updating \( \beta \)

Because \( \beta \) does not occur in \( \log p(w|\mu_w, \Sigma_w) \), we only have to take the derivative of Equation: 2 and we immediately apply the result from Equation: 5:

\[ \frac{d}{d\beta} \sum_{z} p(z|X, w_p, \beta_p) \log p(X, z|w, \beta) \]
\[ = \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n|X_n, w_p, \beta_p) \frac{d}{d\beta} \log p(X_n|\tilde{T}_n(z_n), w, \beta). \]
We obtain the value for \( \frac{d}{d\beta} \log p(X_n | \tilde{T}_n(z_n), w, \beta) \) as follows:

\[
\frac{d}{d\beta} \log p(X_n | z_n, w, \beta) = \sum_{r=1}^{R_n} \sum_{k=1}^{K} \frac{d}{d\beta} \log p(x_{n,r,k} | \tilde{t}_{n,r,k}(z_n), w, \beta)
\]

\[
= \sum_{r=1}^{R_n} \sum_{k=1}^{K} \frac{d}{d\beta} \log N(x_{n,r,k} w | \tilde{t}_{n,r,k}(z_n), \beta^{-1})
\]

\[
= \sum_{r=1}^{R_n} \sum_{k=1}^{K} \frac{1}{2} \frac{d}{d\beta} \log \left( \frac{\beta}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\beta}{2} (x_{n,r,k} w - \tilde{t}_{n,r,k}(z_n))^2 \right)
\]

Plugging this into the previous equation gives us:

\[
\frac{d}{d\beta} \sum_{z} p(z | X, w, \beta_p) \log p(X, z | w, \beta)
\]

\[
= \sum_{n=1}^{2N} \sum_{z_n=1}^{K} p(z_n | X_n, w, \beta_p) \frac{1}{2} \sum_{r=1}^{R_n} \sum_{k=1}^{K} \left( \frac{1}{\beta} - (x_{n,r,k} w - \tilde{t}_{n,r,k}(z_n))^2 \right).
\]

This can then be solved for \( \beta \) after setting it equal to 0. The resulting equation is:

\[
\beta^{-1} = < \sum_{z_n=1}^{K} p(z_n | X_n, w, \beta_p) (x_{n,r,k} w - \tilde{t}_{n,r,k}(z_n))^2 >_{n,r,k}
\]

### 2.3 An update for \( \alpha \)

When \( \mu_w = 0 \) and \( \Sigma_w = \alpha^{-1}I \) it is also possible to optimize \( \alpha \). To optimize \( \alpha \), we only have to consider \( \log p(w | \alpha) \):

\[
\frac{d}{d\alpha} \log N(w | 0, \alpha^{-1}I)
\]

\[
= \frac{d}{d\alpha} \left( \log \frac{1}{\alpha^{-1}I}^{\frac{1}{2}} - \frac{\alpha}{2} w^T w \right)
\]

\[
= D \frac{d}{d\alpha} \log \alpha - \frac{1}{2} w^T w
\]

\[
= D \frac{1}{2} \frac{1}{\alpha} - \frac{1}{2} w^T w.
\]
From which we obtain the final update

$$\alpha = \frac{D}{w^T w}.$$  

In this case the bias term is also regularized. This might be a bad idea but we had much more problems with $\alpha$ going to infinity when the bias was left out of the optimization. The prior distribution on the labels forces the bias in the direction of -1. When the bias is not regularized, then there is nothing forcing the model to obtain a good fit. This results in a weight vector containing all zeros and a single -1 for the bias.