A. Solution of ADM

The underlying algorithm to solve problem (8) is based on the ADM method, (see e.g. Bertsekas & Tsitsiklis, 1989). It consists of iteratively applying the update equations

(a) $\boldsymbol{W}^{[i+1]} \leftarrow \underset{\boldsymbol{W}}{\operatorname{argmin}}\ \mathcal{L}\left(\boldsymbol{W}, \mathbf{C}^{[i]}\boldsymbol{B}^{[i]}\right)$

(b) $\boldsymbol{B}_n^{[i+1]} \leftarrow \underset{\boldsymbol{B}_n}{\operatorname{argmin}}\ \mathcal{L}\left(\boldsymbol{W}^{[i+1]}, \mathbf{C}^{[i]}, \boldsymbol{B}\right)$

(c) $\mathbf{C}_n^{[i+1]} \leftarrow \mathbf{C}_n^{[i]} - \left(\beta \boldsymbol{W}^{[i+1]} - \boldsymbol{B}_n^{[i+1]}\right)$

for $n = 1, \ldots, N$, where $\mathcal{L}$ is the augmented Lagrangian for problem (8) and is defined in equation (7).

We now discuss each of these steps in turn.

Minimizing over $\boldsymbol{W}$

In order to solve Step (a), we need to solve the problem

$$\min_{\boldsymbol{W}} \Big\{ F(\boldsymbol{W}) - \sum_{n=1}^{N} \left(\langle \mathbf{C}_n, \boldsymbol{W} - \boldsymbol{B}_n \rangle + \tfrac{\beta}{2} \left\| \boldsymbol{W} - \boldsymbol{B}_n \right\|_{\mathrm{Fr}}^2 \right) \Big\}$$

which is equal to

$$\min_{\boldsymbol{W}} \Big\{ F(\boldsymbol{W}) - \Big\langle \sum_{n=1}^{N} \mathbf{C}_n + \beta \boldsymbol{B}_n, \boldsymbol{W} \Big\rangle + \frac{N\beta}{2} \left\| \boldsymbol{W} \right\|_{\mathrm{Fr}}^2 + c \Big\},$$

for some constant c whose value is independent of $\boldsymbol{W}$.

Notice that the terms where the whole tensor $\boldsymbol{W}$ appears are both the square of its Frobenius norm and inner products with other tensors. By using the definition of the tensor inner products, it is easy to see that in both cases we can decouple the whole tensor $\boldsymbol{W}$ in terms of the fibers of its mode-1 unfolding, that is the original tasks weight vectors: $\langle \boldsymbol{Z}, \boldsymbol{W} \rangle = \sum_{t=1}^{T} \langle \boldsymbol{Z}_{:,t}, \boldsymbol{W}_{:t} \rangle, \forall \boldsymbol{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_N}$. Consequently, solving the above optimization problem is equivalent to solving the following $T = p_2 p_3 \ldots p_N$ minimization problems

$$\min_{w} \sum_{i=1}^{m_t} L\left(\langle x_i^t, w_t \rangle, y_i^t\right) - \Big\langle \Big(\sum_{n=1}^{N} \mathbf{C}_n + \beta \boldsymbol{B}_n \Big)_{(1),t}, w \Big\rangle + \tfrac{N\beta}{2} \left\| w_t \right\|_{\mathrm{Fr}}^2, \tag{11}$$

for all $t \in \{1, \ldots, T\}$, where we use the notation $w_t = \hat{\boldsymbol{W}}_{(1),t}$. In particular, if we consider one half of the square loss function, then the solution to problem (11) has the close form

$$w_t = \left(X^t X^{t\top} + N\beta I\right)^{-1} \Big[X^t y^t + \Big(\sum_{n=1}^{N} \mathbf{C}_n + \beta \boldsymbol{B}_n \Big)_{(1),t} \Big]$$

where X^t is the $d \times m_t$ data matrix for task t, that is, the columns of X^t are the inputs x_i^t, $i = 1, \ldots, m_t$, and $y^t = (y_1^t, \ldots, y_{m_t}^t)^\top$.

Minimizing over $\boldsymbol{B}_n$

Minimizing equation (7) over $\boldsymbol{B}_n$ is equivalent to the problem

$$\min_{B_{n\,(n)}} \gamma \left\| B_{n\,(n)} \right\|_{1,1} - \langle C_{n\,(n)}, W_{(n)} - B_{n\,(n)} \rangle + \frac{\beta}{2} \left\| W_{(n)} - B_{n\,(n)} \right\|_{\mathrm{Fr}}^2$$

which is the same as the problem

$$\min_{B_{n\,(n)}} \frac{\gamma}{\beta} \left\| B_{n\,(n)} \right\|_{1,1} + \Big\langle \frac{1}{\beta} C_{n(n)} - W_{(n)}, B_{n(n)} \Big\rangle + \frac{1}{2} \left\| B_{n\,(n)} \right\|_{\mathrm{Fr}}^2 + Q_1$$

which in turn equals to

$$\min_{B_{n\,(n)}} \tfrac{\gamma}{\beta} \left\| B_{n\,(n)} \right\|_{1,1} + \tfrac{1}{2} \left\| B_{n\,(n)} - \left(\tfrac{1}{\beta} C_{n(n)} - W_{(n)} \right) \right\|_{\mathrm{Fr}}^2 + Q_2, \tag{12}$$

for some constant matrices $Q_1, Q_2 \in \mathbb{R}^{p_n \times J_n}$. The solution to problem (12) is given by (Gandy et al., 2011) as

$$\hat{B}_{n\,(n)} = \operatorname{shrink}\left(\tfrac{1}{\beta} C_{n(n)} - W_{(n)},\ \tfrac{\gamma}{\beta} \right),$$

where $\operatorname{shrink}(M, k)$ is a function that shrinks the eigenvalues of the matrix M by k. That is, given $M = U \Sigma V^T$, where Σ is a diagonal matrix containing the singular values of M, then $\operatorname{shrink}(M, k) = U S_k(\Sigma) V^T$, where $S_k(\Sigma) = \operatorname{diag}(\max\{\Sigma_{i,i} - k,\ 0\})$.