A Derivation of the f-divergence Lower Bound

A detailed derivation of (6) is given here. The lower bound of the f-divergence is given as

$$\begin{aligned} & d_F(P^T(\boldsymbol{h}), Q^s(\boldsymbol{h})) \\ = & \int P^T(\boldsymbol{h}) F\left(\frac{Q^s(\boldsymbol{h})}{P^T(\boldsymbol{h})}\right) d\boldsymbol{h} \\ \geq & \int Q^s(\boldsymbol{h}) r^s(\boldsymbol{h}) d\boldsymbol{h} - \int P^T(\boldsymbol{h}) \left(\tfrac{r^s(\boldsymbol{h})^2}{2} + r^s(\boldsymbol{h})\right) d\boldsymbol{h} \\ = & -\frac{1}{2}\int P^T(\boldsymbol{h}) r^s(\boldsymbol{h})^2 d\boldsymbol{h} + \int Q^s(\boldsymbol{h}) r^s(\boldsymbol{h}) d\boldsymbol{h} - \\ & \int P^T(\boldsymbol{h}) r_s(\boldsymbol{h}) d\boldsymbol{h} \\ = & -\frac{1}{2}\int P^T(\boldsymbol{h}) \left(\boldsymbol{\alpha}^s \boldsymbol{\phi}^s(\boldsymbol{h})\right)^\intercal \left(\boldsymbol{\alpha}^s \boldsymbol{\phi}^s(\boldsymbol{h})\right) d\boldsymbol{h} \\ & + \int P^s(\boldsymbol{h}|\boldsymbol{y}) \boldsymbol{\gamma}^T (\boldsymbol{\alpha}^s)^\intercal \boldsymbol{\phi}^s(\boldsymbol{h}) d\boldsymbol{h} - 1 \end{aligned}$$

The second step follows the lower bound of f-divergence. The last step is derived because that $r^s(\boldsymbol{h})$ is an estimation of $\frac{Q^s(\boldsymbol{h})}{P^T(\boldsymbol{h})}$. (7) can be derived by re-arranging the terms.

B Proof of Theorem 3.1

In this section, we give the proof sketch of Theorem 3.1. For completeness, we repeat the theorem here.

Theorem. *Assume that* $P^s(\boldsymbol{h}|\boldsymbol{y}) = P^T(\boldsymbol{h}|\boldsymbol{y}) = P(\boldsymbol{h}|\boldsymbol{y})$*, the variance in the feature space is finite, and the label proportions are all non-zero. When the number of training and testing samples goes to infinity,* $\hat{\boldsymbol{\gamma}}^T$ *is asymptotically consistent for* $\boldsymbol{\gamma}^T$ *if* $(\boldsymbol{M}^s)^\intercal \boldsymbol{M}^s$ *is invertible for all* s*.*

Proof. Considering first a single source domain, the quadratic form in Equation 3 would give that the estimator is

$$\hat{\boldsymbol{\gamma}} = ((\boldsymbol{M}^s)^\intercal (\boldsymbol{M}^s))^{-1} (\boldsymbol{M}^s)^\intercal \boldsymbol{\mu}^T. \tag{13}$$

From the assumption that we have finite variance and that conditional distributions are equivalent for all source and target domains, the central limit theorem gives that as the number of samples increases

$$(\boldsymbol{\mu}_l^s - \boldsymbol{\mu}_l^*) \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{n_s \gamma_l^s} \boldsymbol{\Sigma}_l), \tag{14}$$

This is equivalent to noting that $\boldsymbol{M}^s$ is asymptotically centered around $\boldsymbol{M}^*$, because $\boldsymbol{M}^s$ is just the concatenation of these individual vectors. Likewise, we have that asymptotically

$$\boldsymbol{\epsilon} = (\boldsymbol{\mu}^T - \boldsymbol{M}^* \boldsymbol{\gamma}^T) \sim \mathcal{N}(\boldsymbol{0}, \sum_{l=1}^{L} \frac{1}{n_T \gamma_l^T} \boldsymbol{\Sigma}_l). \tag{15}$$

Using the equality from (15) in our estimator of (13), we have that

$$\hat{\boldsymbol{\gamma}} = ((\boldsymbol{M}^s)^\intercal (\boldsymbol{M}^s))^{-1} (\boldsymbol{M}^s)^\intercal (\boldsymbol{M}^* \boldsymbol{\gamma}^T + \boldsymbol{\epsilon}). \tag{16}$$

Note that asymptotically the errors go to 0 from (14) and (15) on all terms in (16), so the estimator has the asymptotic expectation of

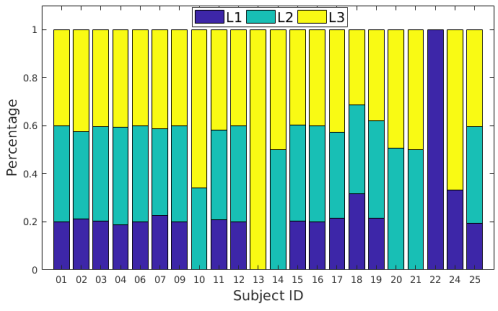
$$\begin{aligned} & \lim_{n_s \to \infty, n_T \to \infty} \mathbb{E}[\hat{\boldsymbol{\gamma}}^T] \\ = & \lim_{n_s \to \infty, n_T \to \infty} \mathbb{E}[((\boldsymbol{M}^s)^\intercal (\boldsymbol{M}^s))^{-1} (\boldsymbol{M}^s)^\intercal (\boldsymbol{M}^* \boldsymbol{\gamma}^T + \epsilon)] \\ = & ((\boldsymbol{M}^*)^\intercal (\boldsymbol{M}^*))^{-1} (\boldsymbol{M}^*)^\intercal \boldsymbol{M}^* \boldsymbol{\gamma}^T = \boldsymbol{\gamma}^T, \end{aligned}$$

if the inverse exists. Therefore, by the weak law of large numbers, $\lim_{n_s \to \infty, n_T \to \infty} \hat{\boldsymbol{\gamma}}^T = \boldsymbol{\gamma}^T$. □

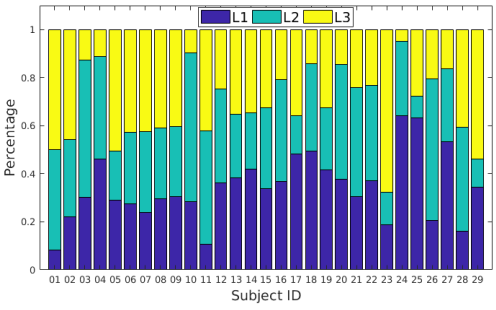
In two domains, $\lim_{n_s \to \infty, n_T \to \infty} \hat{\boldsymbol{\gamma}}^T = \boldsymbol{\gamma}^T$ shows the final result of Theorem 3.1.

C Time Series Data Label Proportion

The label proportion of the two EEG datasets used in experiment Section 5.2 are visualized in Figure 3.



(a) ASD Dataset Label Proportion



(b) LFP Dataset Label Proportion

Figure 3: (Left) Label constitution of the ASD dataset in each domain. 'L1', 'L2', 'L3' means before treatment, six months after treatment and twelve months after treatment, respectively. (Right) Label constitution of the LFP dataset. This dataset is for mice behavior classification using LFP signal. 'L1', 'L2', 'L3' mean home cage, open field and tail suspension, respectively. Please refer the the text for more detailed introduction.