Problem 1: Theorem 4 & 5

Submitted by Amin Shojaeighadikolaei

In this part I want to simulate the first paragraph of the paper which is related to the theorem 4. Theorem 4 is about the theoretical solution for bias. For doing this we need to construct the NN. For doing this based on the paper, we have to generate n random sample with specific distribution. In hence, at the first step I generated 5000 samples in which $|\mathcal{X}| = 16$ and $|\mathcal{Y}| = 8$. I did this in following part of the code:

```
In [4]: # randomly pick joint distribution, normalize
        Pxy = np.random.random([output_size, input_size])
        Pxy = Pxy / np.sum(np.sum(Pxy,axis=0),axis=0)
        #  # compute marginals
        Px = np.sum(Pxy, axis = 0)
        Py = np.sum(Pxy, axis = 1)

        X=np.random.choice(range(input_size),nSamples,p=Px)
        Y=np.random.choice(range(output_size),nSamples,p=Py)

        x_train = One_hot(X)
        y_train = One_hot(Y)
```

As it shows, based on the experimental validation part, input and output should be one hot encoded.

Then canonical dependence matrix is needed which is calculate by following formula:

$$\tilde{B}(x,y) \triangleq \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}} \tag{1}$$

Based on **lemma 1**, the SVD of $\tilde{B}$ can be written as $\tilde{B} = \sum_{i=1}^{K} \sigma_i \psi_i^Y (\psi_i^X)^T$ in which σ_i denotes the ith singular value with the ordering left and right singular vectors. So based on the paper, the optimal features are:

$$f_i(x) = \psi_i^X / \sqrt{P_X(x)} \tag{2}$$
$$g_i(y) = \psi_i^Y / \sqrt{P_Y(y)} \tag{3}$$

These optimal features are proposed in following part of the code:

```
In [5]: B = (Pxy- Px.reshape(1, -1)*Py.reshape(-1, 1))/ np.sqrt(Px.reshape(1, -1)) * np.sqrt(Py.reshape(-1, 1))
        phi_y, phi, phi_x = np.linalg.svd(B)
        f_i = phi_x[1,:] / np.sqrt(Px)
        g_i = phi_y[:,1] / np.sqrt(Py)
```

The next part is constructing the NN with 1 hidden layer(with one node).

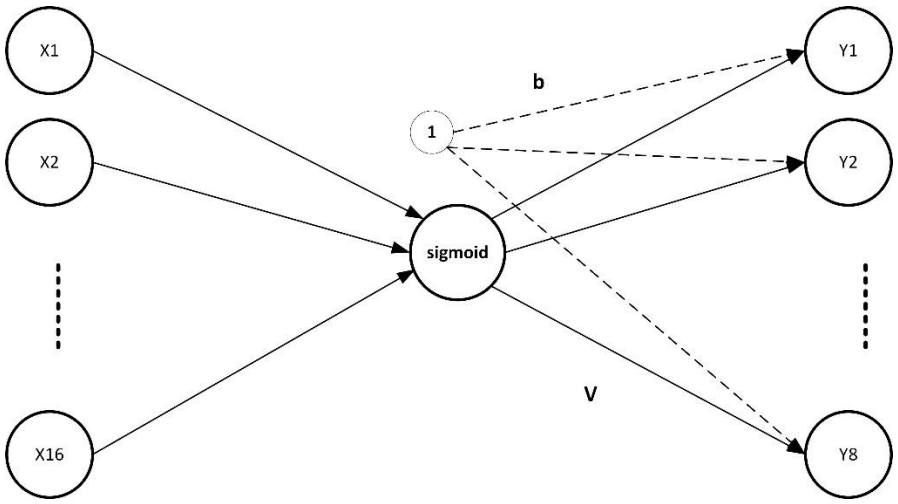


Fig1: Neural Network for theorem 4

Based on theorem 4, the optimal solution for weight and bias are given by :

$$\tilde{d}(y) = -\mu_s^T \tilde{v}(y) \tag{4}$$

In addition, from K-L divergence we know that:

$$d(y) = b(y) - \log P_Y(y) \ , for\ y \epsilon \mathcal{Y} \tag{5}$$

From (4) and (5) we have:

$$b(y) = \log P_Y(y) - \ \mu_s^T \tilde{v}(y) \tag{6}$$

In which $\mu_s = \mathbb{E}[X]$

The result is shown by Fig2 that training result is match to the theoretical result.

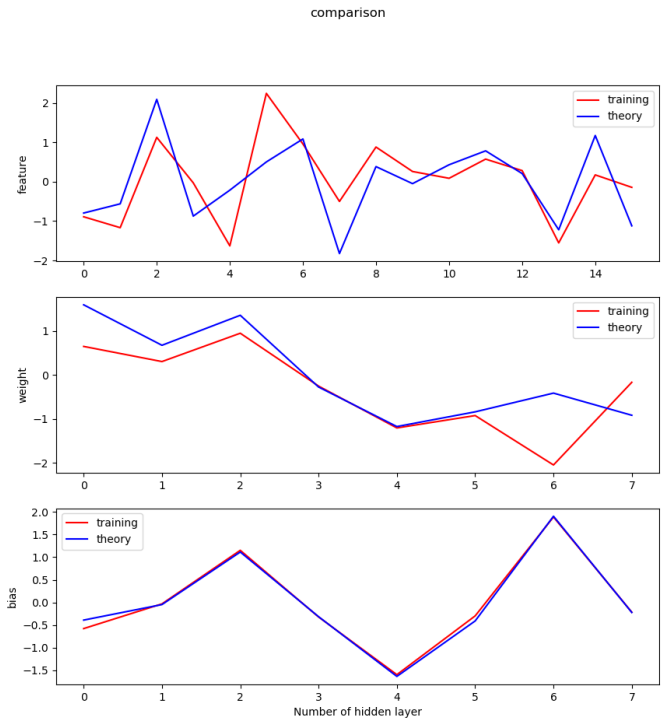
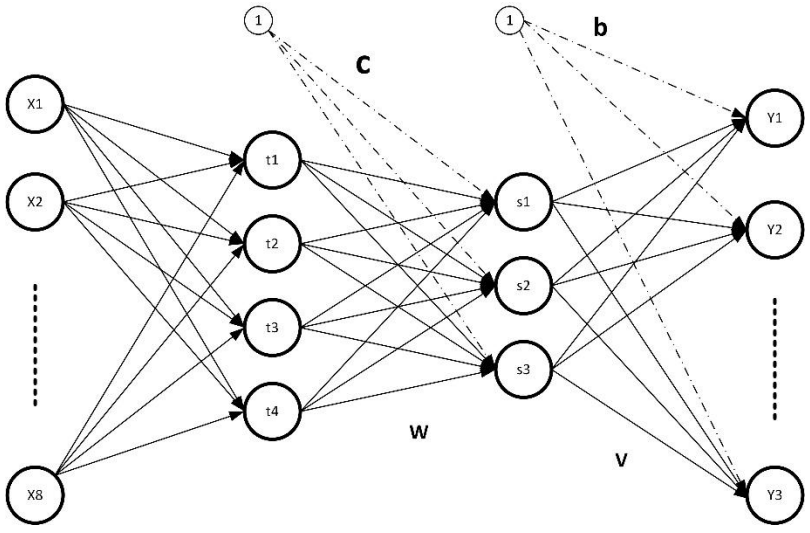


Fig2: comparison between theory and training

For theorem 5 same procedure is considered for creating the samples but here we have 2 hidden layers same as following:



From the Lemma 2. We know that the information vectors are:

$$\xi^X(x) = \sqrt{P_X(x)}\tilde{s}(x) \quad (7)$$

$$\xi^Y(y) = \sqrt{P_Y(y)}\tilde{v}(y) \quad (8)$$

Bias for the final layer (b) can be calculated as (6) but since here we have 2 hidden layers the for μ_s we have:

$$\mu_s = \mathbb{E}[s] \quad (9)$$