A. Proof of Theorem 1

Proof. First of all,

\[
\mathbb{P}(X, Y = \bar{y}) = \frac{1}{K - 1} \sum_{y \neq \bar{y}} \mathbb{P}(X, Y = y)
\]

\[
= \frac{1}{K - 1} \left( \sum_{y=1}^{K} \mathbb{P}(X, Y = y) - \mathbb{P}(X, Y = \bar{y}) \right)
\]

\[
= \frac{1}{K - 1} \left( \mathbb{P}(X) - \mathbb{P}(X, Y = \bar{y}) \right).
\]

The first equality holds since the marginal distribution is equivalent for \(D\) and \(\overline{D}\) and we assume (5). Consequently,

\[
\mathbb{P}(Y = \bar{y} | X = x) = \frac{\mathbb{P}(X = x, Y = \bar{y})}{\mathbb{P}(X = x)}
\]

\[
= \frac{1}{K - 1} \cdot \left( 1 - \frac{\mathbb{P}(X, Y = \bar{y})}{\mathbb{P}(X = x)} \right)
\]

\[
= \frac{1}{K - 1} \cdot \left( 1 - \mathbb{P}(Y = \bar{y} | X = x) \right)
\]

\[
= -\frac{1}{K - 1} \mathbb{P}(Y = \bar{y} | X = x) + \frac{1}{K - 1}.
\]

More simply, we have \(\eta(x) = -(K - 1)\eta(x) + 1\). Finally, we transform the classification risk,

\[
R(g; \ell) = \mathbb{E}_{(X,Y)\sim D}(\ell(Y, g(X))] = \mathbb{E}_{X\sim M}[\eta^\top \ell(g(X))]
\]

\[
= \mathbb{E}_{X\sim M} \left[ (-(K-1)\eta^\top + 1^\top) \ell(g(X)) \right]
\]

\[
= \mathbb{E}_{X\sim M} \left[ -(K-1)\eta^\top \ell(g(X)) + 1^\top \ell(g(X)) \right]
\]

\[
= \mathbb{E}_{(X,Y)\sim \mathcal{P}} \left[ -(K-1) \cdot \ell(Y, g(X)) \right] + 1^\top \mathbb{E}_{X\sim M} [\ell(g(X))]
\]

\[
= \sum_{k=1}^{K} \mathbb{P}_k \cdot \mathbb{E}_{X\sim \mathcal{P}_k} \left[ -(K-1) \cdot \ell(k, g(X)) \right]
\]

\[
+ 1^\top \mathbb{E}_{X\sim M} [\ell(g(X))]
\]

\[
= \overline{\ell}(g; \ell)
\]

for the complementary loss, \(\overline{\ell}(k, g) := -(K-1)\ell(k, g) + 1^\top \ell(g)\), which concludes the proof. \(\square\)

B. Proof of Corollary 2

Proof.

\[
\overline{\ell}(g; \ell) = \mathbb{E}_{\mathcal{P}}[\overline{\ell}(Y, g(X))]
\]

\[
= \mathbb{E}_{\mathcal{P}}[-(K-1)\ell(Y, g(X)) + \sum_{j=1}^{K} \ell(j, g(X))]
\]

\[
= \mathbb{E}_{\mathcal{P}}[(-K-1)[M_2 - \overline{\ell}(Y, g(X))] + M_1]
\]

\[
= (K-1)\mathbb{E}_{\mathcal{P}}[\overline{\ell}(Y, g(X))] + M_1 - (K-1)M_2
\]

\[
= (K-1)\mathbb{E}_{\mathcal{P}}[\overline{\ell}(Y, g(X))] - M_1 + M_2
\]

Table 3: Summary statistics of benchmark datasets. In the experiments with validation dataset in Section 4.2, train data is further split into train/validation with a ratio of 9:1. Fashion is Fashion-MNIST and Kuzushi is Kuzushi-MNIST.

<table>
<thead>
<tr>
<th>Name</th>
<th># Train</th>
<th># Test</th>
<th># Dim</th>
<th># Classes</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>60k</td>
<td>10k</td>
<td>784</td>
<td>10</td>
<td>Linear, MLP</td>
</tr>
<tr>
<td>Fashion</td>
<td>60k</td>
<td>10k</td>
<td>784</td>
<td>10</td>
<td>Linear, MLP</td>
</tr>
<tr>
<td>Kuzushi</td>
<td>60k</td>
<td>10k</td>
<td>784</td>
<td>10</td>
<td>Linear, MLP</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>50k</td>
<td>10k</td>
<td>2,048</td>
<td>10</td>
<td>DenseNet, ResNet</td>
</tr>
</tbody>
</table>

The second equality holds because we use (10). The third equality holds because we are using losses that satisfy \(\sum_j \ell_j(g(x)) = M_1\) for all \(x\) and \(\ell(\overline{y}, g(x)) + \overline{\ell}(\overline{y}, g(x)) = M_2\) for all \(x\) and \(\overline{y}\). The fourth equality rearranges terms. The 5th equality holds because \(M_1 - (K-1)M_2 = -M_1 + M_2\) for \(\overline{\ell}_{OVA}\) and \(\overline{\ell}_{PC}\). This can be easily shown by using \(M_1 = K\) and \(M_2 = 2\) for \(\overline{\ell}_{OVA}\), and \(M_1 = K(K-1)/2\) and \(M_2 = K - 1\) for \(\overline{\ell}_{PC}\). \(\square\)

C. Datasets

In the experiments in Section 4, we use 4 benchmark datasets explained below. The summary statistics of the four datasets are given in Table 3.

- MNIST\(^4\) (Lecun et al., 1998) is a 10 class dataset of handwritten digits: 1, 2, \ldots, 9 and 0. Each sample is a 28 \times 28 grayscale image.
- Fashion-MNIST\(^5\) (Xiao et al., 2017) is a 10 class dataset of fashion items: T-shirt/top, Trouser, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. Each sample is a 28 \times 28 grayscale image.
- Kuzushi-MNIST\(^6\) (Clanuwat et al., 2018) is a 10 class dataset of cursive Japanese (“Kuzushiji”) characters. Each sample is a 28 \times 28 grayscale image.
- CIFAR-10\(^7\) is a 10 class dataset of various objects: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each sample is a colored image in 32 \times 32 \times 3 RGB format. It is a subset of the 80 million tiny images dataset (Torralba et al., 2008).

\(^4\)http://yann.lecun.com/exdb/mnist/
\(^5\)https://github.com/zalandoresearch/fashion-mnist
\(^6\)https://github.com/rois-codh/kmnist
\(^7\)https://www.cs.toronto.edu/~kriz/cifar.html