A. Proof of Theorem 1

Proof. First of all,

$$\begin{aligned}\mathbb{P}(X, \overline{Y} = \overline{y}) &= \frac{1}{K-1}\sum_{y \neq \overline{y}} \mathbb{P}(X, Y = y) \\ &= \frac{1}{K-1}\Big(\sum_{y=1}^{K} \mathbb{P}(X, Y = y) - \mathbb{P}(X, Y = \overline{y})\Big) \\ &= \frac{1}{K-1}\big(\mathbb{P}(X) - \mathbb{P}(X, Y = \overline{y})\big).\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(\overline{Y} = \overline{y} | X = x) &= \frac{\mathbb{P}(X = x, \overline{Y} = \overline{y})}{\mathbb{P}(X = x)} \\ &= \frac{1}{K-1}\cdot\Big(1 - \frac{\mathbb{P}(X, Y = \overline{y})}{\mathbb{P}(X = x)}\Big) \\ &= \frac{1}{K-1}\cdot\big(1 - \mathbb{P}(Y = \overline{y} | X = x)\big) \\ &= -\frac{1}{K-1}\mathbb{P}(Y = \overline{y} | X = x) + \frac{1}{K-1}.\end{aligned}$$

More simply, we have $\boldsymbol{\eta}(x) = -(K-1)\overline{\boldsymbol{\eta}}(x) + \mathbf{1}$. Finally, we transform the classification risk,

$$\begin{aligned}R(g; \ell) &= \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(Y, \boldsymbol{g}(X))] = \mathbb{E}_{X\sim M}[\boldsymbol{\eta}^\top \boldsymbol{\ell}(\boldsymbol{g}(X))] \\ &= \mathbb{E}_{X\sim M}\big[\big(-(K-1)\overline{\boldsymbol{\eta}}^\top + \mathbf{1}^\top\big)\boldsymbol{\ell}\big(\boldsymbol{g}(X)\big)\big] \\ &= \mathbb{E}_{X\sim M}\big[-(K-1)\overline{\boldsymbol{\eta}}^\top \boldsymbol{\ell}\big(\boldsymbol{g}(X)\big) + \mathbf{1}^\top \boldsymbol{\ell}\big(\boldsymbol{g}(X)\big)\big] \\ &= \mathbb{E}_{(X,\overline{Y})\sim\overline{\mathcal{D}}}\big[-(K-1)\cdot\ell\big(\overline{Y}, \boldsymbol{g}(X)\big)\big] \\ &\qquad + \mathbf{1}^\top \mathbb{E}_{X\sim M}\big[\boldsymbol{\ell}\big(\boldsymbol{g}(X)\big)\big] \\ &= \sum_{k=1}^{K} \overline{\pi}_k \cdot \mathbb{E}_{X\sim\overline{P}_k}\Big[-(K-1)\cdot\ell\big(k, \boldsymbol{g}(X)\big) \\ &\qquad + \mathbf{1}^\top \boldsymbol{\ell}\big(\boldsymbol{g}(X)\big)\Big] \\ &= \overline{R}(g; \overline{\ell})\end{aligned}$$

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

B. Proof of Corollary 2

Proof.

$$\begin{aligned}\overline{R}(\boldsymbol{g}; \overline{\ell}) &= \mathbb{E}_{\overline{\mathcal{D}}}[\overline{\ell}(\overline{Y}, \boldsymbol{g}(X))] \\ &= \mathbb{E}_{\overline{\mathcal{D}}}[-(K-1)\ell(\overline{Y}, \boldsymbol{g}(X)) + \sum_{j=1}^{K} \ell(j, \boldsymbol{g}(X))] \\ &= \mathbb{E}_{\overline{\mathcal{D}}}\big[-(K-1)[M_2 - \overline{\ell}(\overline{Y}, \boldsymbol{g}(X))] + M_1\big] \\ &= (K-1)\mathbb{E}_{\overline{\mathcal{D}}}[\overline{\ell}(\overline{Y}, \boldsymbol{g}(X))] + M_1 - (K-1)M_2 \\ &= (K-1)\mathbb{E}_{\overline{\mathcal{D}}}[\overline{\ell}(\overline{Y}, \boldsymbol{g}(X))] - M_1 + M_2\end{aligned}$$

The second equality holds because we use (10). The third equality holds because we are using losses that satisfy $\sum_j \ell(j, \boldsymbol{g}(x)) = M_1$ for all x and $\ell(\overline{y}, \boldsymbol{g}(x)) + \overline{\ell}(\overline{y}, \boldsymbol{g}(x)) = M_2$ for all x and $\overline{y}$. The 4th equality rearranges terms. The 5th equality holds because $M_1 - (K-1)M_2 = -M_1 + M_2$ for $\overline{\ell}_{\mathrm{OVA}}$ and $\overline{\ell}_{\mathrm{PC}}$. This can be easily shown by using $M_1 = K$ and $M_2 = 2$ for $\overline{\ell}_{\mathrm{OVA}}$, and $M_1 = K(K-1)/2$ and $M_2 = K-1$ for $\overline{\ell}_{\mathrm{PC}}$. □

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

Name	# Train	# Test	# Dim	# Classes	Model
MNIST	60k	10k	784	10	Linear, MLP
Fashion	60k	10k	784	10	Linear, MLP
Kuzushi	60k	10k	784	10	Linear, MLP
CIFAR-10	50k	10k	2,048	10	DenseNet, Resnet

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 28×28 grayscale image.
- Fashion-MNIST[5] (Xiao et al., 2017) is a 10 class dataset of fashion items: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. Each sample is a 28×28 grayscale image.
- Kuzushi-MNIST[6] (Clanuwat et al., 2018) is a 10 class dataset of cursive Japanese ("Kuzushiji") characters. Each sample is a 28×28 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