Appendix A. Proofs of Loss Function Properties

As stated in Section 4, if we define a loss function $\boldsymbol{\mathcal{L}}$ in an additive form by (12), there exist constants ξ_1 and ξ_2 which satisfy (10) and (11). We first describe proof of (10). The following formulation holds in accordance with $\ell(z) + \ell(-z) = a$.

$$
\begin{aligned}
&\sum_{y \in Y} \mathcal{L}_{\mathrm{OVA}}(f(x), y) \\
&= \sum_{y \in Y} \left(\ell(g_y(x)) + \frac{1}{K-1} \sum_{y' \neq y} \ell\left(-g_{y'}(x)\right) \right) \\
&= \sum_{y \in Y} \left(\frac{K}{K-1} \ell(g_y(x)) - \frac{1}{K-1} \ell(g_y(x)) + \frac{1}{K-1} \sum_{y' \in \mathcal{Y}} \ell\left(-g_{y'}(x)\right) - \frac{1}{K-1} \ell(-g_y(x)) \right) \\
&= \frac{1}{K-1} \sum_{y \in Y} \left(K \ell(g_y(x)) + \sum_{y' \in \mathcal{Y}} \ell\left(-g_{y'}(x)\right) - a \right) \\
&= \frac{1}{K-1} \left(K \sum_{y \in Y} \ell(g_y(x)) + \sum_{y \in Y} \sum_{y' \in \mathcal{Y}} \ell\left(-g_{y'}(x)\right) - \sum_{y \in Y} a \right) \\
&= \frac{1}{K-1} \left((K - N + N) \sum_{y \in Y} \ell(g_y(x)) + N \sum_{y \in Y} \ell(-g_y(x)) + N \sum_{y' \notin Y} \ell\left(-g_{y'}(x)\right) - aN \right) \\
&= \frac{1}{K-1} \left((K - N) \sum_{y \in Y} \ell(g_y(x)) + N \sum_{y' \notin Y} \ell\left(-g_{y'}(x)\right) + aN(N-1) \right) \\
&= \frac{K-N}{K-1} \sum_{y \in Y} \ell(g_y(x)) + \frac{N}{K-1} \sum_{y' \notin Y} \ell\left(-g_{y'}(x)\right) + \frac{aN(N-1)}{K-1} \\
&= \boldsymbol{\mathcal{L}}_{\mathrm{OVA}}(f(x), Y) + \frac{aN(N-1)}{K-1}
\end{aligned}
$$

■

Next, we describe proof of (11). Similar to the above formulation, in accordance with $\ell(z) + \ell(-z) = a$,

$$
\begin{aligned}
&\sum_{y \in Y} \mathcal{L}_{\mathrm{PC}}(f(x), y) \\
&= \sum_{y \in Y} \sum_{y' \neq y} \ell\left(g_y(x) - g_{y'}(x)\right) \\
&= \sum_{y \in Y} \sum_{y' \in \mathcal{Y}} \ell\left(g_y(x) - g_{y'}(x)\right) - \sum_{y \in Y} \ell(g_y(x) - g_y(x))
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{y\in Y}\sum_{y'\notin Y} \ell\left(g_y(x) - g_{y'}(x)\right) + \sum_{\substack{y,y'\in Y\\ y=y'}} \ell\left(g_y(x) - g_{y'}(x)\right) + \sum_{\substack{y,y'\in Y\\ y\neq y'}} \ell\left(g_y(x) - g_{y'}(x)\right) - \frac{aN}{2} \\
&= \sum_{y\in Y}\sum_{y'\notin Y} \ell\left(g_y(x) - g_{y'}(x)\right) + a \cdot {}_N C_2 \\
&= \boldsymbol{\mathcal{L}}_{\mathrm{PC}}\left(f(x), Y\right) + a \cdot {}_N C_2
\end{aligned}
$$

The third and fourth equality hold due to $\ell\left(0\right) = a/2$. ■

Appendix B. Property of the Data-Generation Probability Model

In this section, we describe proof of (15). To begin with, we introduce the following lemma.

Lemma 5 *Let any finite sets be X with size of K, and any elements of the N-size power set $\mathfrak{P}_N\left(X\right)$ be A. Then, the following equation holds for any function f and g over X.*

$$
\sum_{A\in\mathfrak{P}_N(X)}\sum_{a_1,a_2\in A} f(a_1)g(a_2) = {}_{K-2}C_{N-2}\sum_{\substack{x_1,x_2\in X\\ x_1\neq x_2}} f(x_1)g(x_2) + {}_{K-1}C_{N-1}\sum_{\substack{x_1,x_2\in X\\ x_1=x_2}} f(x_1)g(x_2) \tag{23}
$$

Proof The left-hand side can be formulated as:

$$
\sum_{A\in\mathfrak{P}_N(X)}\sum_{a_1,a_2\in A} f(a_1)g(a_2) = \sum_{A\in\mathfrak{P}_N(X)}\sum_{\substack{a_1,a_2\in A\\ a_1\neq a_2}} f(a_1)g(a_2) + \sum_{A\in\mathfrak{P}_N(X)}\sum_{\substack{a_1,a_2\in A\\ a_1=a_2}} f(a_1)g(a_2)
$$

For the first term, any A can be chosen from $\mathfrak{P}_N\left(X\right)$ in ${}_K C_N$ patterns, and any $a_1,\ a_2\ (a_1 \neq a_2)$ can be chosen from A in ${}_N P_2$ patterns. Because $f(a_1)g(a_2)$ with any $a_1,\ a_2\ (a_1 \neq a_2)$ can be chosen from X in ${}_K P_2$ patterns, the first term in (23) holds due to ${}_{K-2}C_{N-2} = \frac{{}_K C_N \cdot {}_N P_2}{{}_K P_2}$. Similar for the second term. ■

Here, we prove (15). Let $P(\alpha \in Y|x)$ be the probability where a set of candidate labels Y includes a label $\alpha \in \mathcal{Y}$. Denoting I as an indicator function, then

$$
\begin{aligned}
P(\alpha \in Y) &= \sum_{Y\in\mathfrak{P}_N(\mathcal{Y})} I(\alpha \in Y) P_N\left(Y|x\right) \\
&= \frac{1}{{}_{K-1}C_{N-1}} \sum_{Y\in\mathfrak{P}_N(\mathcal{Y})} I(\alpha \in Y) \sum_{y\in Y} P(y|x) \\
&= \frac{1}{{}_{K-1}C_{N-1}} \sum_{Y\in\mathfrak{P}_N(\mathcal{Y})} \sum_{y_1,y_2\in Y} I(y_1 = \alpha) P(y_2|x) \\
&= \frac{1}{{}_{K-1}C_{N-1}} \left({}_{K-1}C_{N-1} \sum_{y\in\mathcal{Y}} I(y = \alpha) P(y|x) + {}_{K-2}C_{N-2} \sum_{\substack{y_1,y_2\in\mathcal{Y}\\ y_1\neq y_2}} I(y_1 = \alpha) P(y_2|x) \right)
\end{aligned}
$$

$$= P(\alpha|x) + \frac{N-1}{K-1}\sum_{y\neq\alpha} P(y|x)$$
$$= \frac{K-N}{K-1}P(\alpha|x) + \frac{N-1}{K-1}$$

The fourth equality holds due to lemma 5. In the fifth equality, the second term holds because $y_2 \neq \alpha$ holds if $y_1 = \alpha$. Therefore,

$$\begin{aligned} Q(y_o = \alpha|x) &= Q(y_o = \alpha|x, \alpha \in Y)Q(\alpha \in Y|x) + Q(y_o = \alpha|x, \alpha \notin Y)Q(\alpha \notin Y|x) \\ &= \frac{1}{N}\frac{K-N}{N-1}P(\alpha|x) + \frac{1}{N}\frac{N-1}{K-1} \\ &= \beta P(\alpha|x) + (1-\beta)\frac{1}{K} \end{aligned}$$

The second equality holds due to the assumption where $Q(y_o = \alpha|x, \alpha \in Y) = 1/N$ and $Q(y_o = \alpha|x, \alpha \notin Y) = 0$ hold. ■

Appendix C. Proof of Theorem 1

We first derive the expectation of the sum of loss for all the ordinary labels.

$$\begin{aligned} &\mathbb{E}_{P_N(Y|x)}\left[\sum_{y\in Y}\mathcal{L}\left(f(x), y\right)\right] \\ &= \sum_{Y\in\mathfrak{P}_N(\mathcal{Y})}\sum_{y\in Y}\mathcal{L}\left(f(x), y\right)P_N\left(Y|x\right) \\ &= \frac{1}{{}_{K-1}C_{N-1}}\sum_{Y\in\mathfrak{P}_N(\mathcal{Y})}\sum_{y\in Y}\sum_{y'\in Y}\mathcal{L}\left(f(x), y'\right)P_N\left(y|x\right) \\ &= \frac{1}{{}_{K-1}C_{N-1}}\left\{{}_{K-2}C_{N-2}\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\mathcal{L}\left(f(x), y'\right)P_N\left(y|x\right) + {}_{K-1}C_{N-1}\sum_{y\in\mathcal{Y}}\mathcal{L}\left(f(x), y\right)P_N\left(y|x\right)\right\} \\ &= \frac{N-1}{K-1}\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\mathcal{L}\left(f(x), y'\right)P_N\left(y|x\right) + \sum_{y\in\mathcal{Y}}\mathcal{L}\left(f(x), y\right)P_N\left(y|x\right) \\ &= \frac{N-1}{K-1}\mathbb{E}_{P(y|x)}\left[\sum_{y'\in\mathcal{Y}}\mathcal{L}\left(f(x), y'\right) - \mathcal{L}\left(f(x), y\right)\right] + \mathbb{E}_{P(y|x)}\left[\mathcal{L}\left(f(x), y\right)\right] \\ &= \frac{N-1}{K-1}\sum_{y\in\mathcal{Y}}\mathcal{L}\left(f(x), y\right) + \frac{K-N}{K-1}\mathbb{E}_{P(y|x)}\left[\mathcal{L}\left(f(x), y\right)\right] \end{aligned}$$

The second equality holds due to the definition of $P_N\left(Y|x\right)$. The third equality holds due to Lemma 5. Thus, the following formulation holds due to (12).

$$\mathbb{E}_{P_N(x,Y)}\left[\boldsymbol{\mathcal{L}}\left(f(x), Y\right)\right] = \xi_1\mathbb{E}_{P_N(x,Y)}\left[\sum_{y\in Y}\mathcal{L}\left(f(x), y\right)\right] + \xi_2$$

$$= \frac{\xi_1(K-N)}{K-1}\mathbb{E}_{P(x,y)}\left[\mathcal{L}\left(f(x),y\right)\right] + \frac{\xi_1(N-1)}{K-1}\sum_{y\in\mathcal{Y}}\mathcal{L}\left(f(x),y\right) + \xi_2$$

■

Appendix D. Proof of Lemma 2

According to the duality described in (12) and (13), loss function $\boldsymbol{\mathcal{L}}$ can be formulated by $\mathcal{L}\left(f(x),y\right)$ for N candidate labels $y \in Y$, or $\overline{\mathcal{L}}\left(f(x),\overline{y}\right)$ for $K-N$ complementary labels $\overline{y} \in \overline{Y}$. Thus, we can redefine loss function $\boldsymbol{\mathcal{L}}$ as the following:

$$\boldsymbol{\mathcal{L}}\left(f(x),Y\right) = \sum_{y\in\widetilde{Y}}\widetilde{\mathcal{L}}\left(f(x),y\right) = \begin{cases} \sum\limits_{y\in Y}\mathcal{L}\left(f(x),y\right), & \text{if } N \leq \dfrac{K}{2} \\ \sum\limits_{\overline{y}\in\overline{Y}}\overline{\mathcal{L}}\left(f(x),\overline{y}\right), & \text{otherwise} \end{cases}$$

where $\widetilde{\mathcal{L}}$ and $\widetilde{Y}$ denote $\mathcal{L}$ and Y respectively if $N \leq K/2$, otherwise $\overline{\mathcal{L}}$, $\overline{Y}$ respectively. Therefore, given $\widetilde{N} := |\widetilde{Y}|$ it always satisfies $\widetilde{N} \leq K/2$. Note that $\xi_2 = \overline{\xi}_2 = 0$ due to the assumption of $a = 0$, as discussed in Section 4.1. Similarly, $\widetilde{\ell}\left(z\right)$ denotes $\widetilde{\ell} : z \mapsto \ell\left(z\right)$ if $N \leq K/2$ otherwise $\widetilde{\ell} : z \mapsto \ell\left(-z\right)$. For the rest of this work, we prove Lemma 2 according to those definitions.

First we describe proof for one-versus-all classification. Under the assumption of $a = 0$, the following formulation holds.

$$\sum_{y\in\widetilde{Y}}\widetilde{\mathcal{L}}_{\mathrm{OVA}}\left(f(x),y\right) = \sum_{y\in\widetilde{Y}}\left(\frac{K}{K-1}\widetilde{\ell}\left(g_y(x)\right) + \frac{1}{K-1}\sum_{y'\in\mathcal{Y}}\widetilde{\ell}\left(-g_{y'}(x)\right)\right)$$

Thus, the following formulation holds for any $\widetilde{Y}, \widetilde{Y}' \in \mathfrak{P}_N\left(\mathcal{Y}\right)$.

$$\begin{aligned}
\|\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}\|_\infty &= \sup_{g_1,\cdots,g_K\in\mathcal{G}}\left(\sum_{y\in\widetilde{Y}}\widetilde{\mathcal{L}}_{\mathrm{OVA}}\left(f(x),y\right) - \sum_{y'\in\widetilde{Y}'}\widetilde{\mathcal{L}}_{\mathrm{OVA}}\left(f(x),y'\right)\right) \\
&= \sup_{g_1,\cdots,g_K\in\mathcal{G}}\left\{\frac{K}{K-1}\left(\sum_{y\in\widetilde{Y}}\widetilde{\ell}\left(g_y(x)\right) - \sum_{y'\in\widetilde{Y}'}\widetilde{\ell}\left(g_{y'}(x)\right)\right)\right. \\
&\quad \left. + \frac{1}{K-1}\left(\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(-g_y(x)\right) - \sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(-g_y(x)\right)\right)\right\} \\
&\leq \frac{K}{K-1}\left(\sup_{g_1,\cdots,g_K\in\mathcal{G}}\sum_{y\in\widetilde{Y}}\widetilde{\ell}\left(g_y(x)\right) - \inf_{g_1,\cdots,g_K}\sum_{y'\in\widetilde{Y}'}\widetilde{\ell}\left(g_{y'}(x)\right)\right) \\
&\leq \frac{K}{K-1}\left(\frac{\widetilde{N}}{2} + \frac{\widetilde{N}}{2}\right)
\end{aligned}$$

$$= \frac{K\widetilde{N}}{K-1}$$

The second inequality holds because supremum and infimum of ℓ are $1/2$ and $-1/2$ respectively. ■

We further describe proof for pairwise classification. Under the assumption of $a = 0$, the following formulation holds.

$$\sum_{y\in\widetilde{Y}} \widetilde{\mathcal{L}}_{\mathrm{PC}}\left(f(x), y\right) = \sum_{y\in\widetilde{Y}} \sum_{y'\notin\widetilde{Y}} \widetilde{\ell}\left(g_y(x) - g_{y'}(x)\right)$$

Thus,

$$\begin{aligned}
\|\boldsymbol{\mathcal{L}}_{\mathrm{PC}}\|_\infty &= \sup_{g_1,\cdots,g_K\in\mathcal{G}} \left(\sum_{y'\in\widetilde{Y}} \widetilde{\mathcal{L}}_{\mathrm{PC}}\left(f(x), y'\right) - \sum_{y\in\widetilde{Y}'} \widetilde{\mathcal{L}}_{\mathrm{PC}}\left(f(x), y\right) \right) \\
&= \sup_{g_1,\cdots,g_K\in\mathcal{G}} \sum_{y\in\widetilde{Y}} \sum_{y'\notin\widetilde{Y}} \widetilde{\ell}\left(g_y(x) - g_{y'}(x)\right) - \inf_{g_1,\cdots,g_K\in\mathcal{G}} \sum_{y'\in\widetilde{Y}'} \sum_{y\notin\widetilde{Y}'} \widetilde{\ell}\left(g_{y'}(x) - g_y(x)\right) \\
&\leq \frac{\widetilde{N}(K-\widetilde{N})}{2} - \left(-\frac{\widetilde{N}(K-\widetilde{N})}{2}\right) \\
&= \widetilde{N}(K-\widetilde{N})
\end{aligned}$$

■

Appendix E. Proof of Lemma 3

We decribe proof for one-versus-all classification. Under the assumption that $h \in \mathcal{H}_{\mathrm{OVA}}$ is equivalent to $\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}$, the following formulation holds due to the definition of $\mathcal{H}_{\mathrm{OVA}}$.

$$\begin{aligned}
\mathfrak{R}_n\left(\mathcal{H}_{\mathrm{OVA}}\right) &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}_{\mathrm{OVA}}} \frac{1}{n} \sum_{i=1}^n \sigma_i h\left(x_i, Y_i\right) \right] \\
&= \mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma \left[\sup_{g_1,\cdots,g_K\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{y\in\widetilde{Y}_i} \widetilde{\mathcal{L}}_{\mathrm{OVA}}\left(f(x_i), y\right) \right] \\
&= \mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma \left[\sup_{g_1,\cdots,g_K\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{y\in\widetilde{Y}_i} \left\{ \frac{K}{K-1}\widetilde{\ell}\left(g_y(x_i)\right) + \frac{1}{K-1} \sum_{y'\in\mathcal{Y}} \widetilde{\ell}\left(-g_{y'}(x_i)\right) \right\} \right] \\
&\leq \frac{K}{K-1}\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma \left[\sup_{g_1,\cdots,g_K\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{y\in\widetilde{Y}_i} \widetilde{\ell}\left(g_y(x_i)\right) \right]
\end{aligned}$$

$$+\frac{1}{K-1}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\widetilde{Y}_i}\sum_{y'\in\mathcal{Y}}\widetilde{\ell}\left(-g_{y'}(x_i)\right)\right]$$

Let $I(y\in\widetilde{Y}_i)$ be an indicator function and define $\alpha_i:=2I(y\in\widetilde{Y}_i)-1$, then for the first term,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\widetilde{Y}_i}\widetilde{\ell}\left(g_y(x_i)\right)\right]\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(g_y(x_i)\right)I(y\in\widetilde{Y}_i)\right]\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(g_y(x_i)\right)(\alpha_i+1)\right]\\
&\le\sum_{y\in\mathcal{Y}}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(g_y(x_i)\right)(\alpha_i+1)\right]\\
&\le\sum_{y\in\mathcal{Y}}\left\{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\alpha_i\sigma_i\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(g_y(x_i)\right)\right]+\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(g_y(x_i)\right)\right]\right\}\\
&=K\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\mathcal{Y}}\widetilde{\ell}\left(g(x_i)\right)\right]\\
&=K\mathfrak{R}_n\left(\widetilde{\ell}\circ\mathcal{G}\right)
\end{aligned}$$

The second equality from the last holds because σ_i and $\alpha_i\sigma_i$ are drawn from the same probabilistic distribution. For the second term,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\widetilde{Y}_i}\sum_{y'\in\mathcal{Y}}\widetilde{\ell}\left(-g_{y'}(x_i)\right)\right]&\le\sum_{y\in\mathcal{Y}}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_y\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{N}\widetilde{\ell}\left(-g_y(x_i)\right)\right]\\
&=K\widetilde{N}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(-g(x_i)\right)\right]\\
&=K\widetilde{N}\mathfrak{R}_n\left(\widetilde{\ell}\circ\mathcal{G}\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathfrak{R}_n\left(\mathcal{H}_{\mathrm{OVA}}\right)&\le\frac{K^2+K\widetilde{N}}{K-1}\mathfrak{R}_n\left(\widetilde{\ell}\circ\mathcal{G}\right)\\
&\le\frac{K(K+\widetilde{N})}{K-1}L_{\ell}\mathfrak{R}_n\left(\mathcal{G}\right)
\end{aligned}$$

The second inequality holds due to $\mathfrak{R}_n\left(\widetilde{\ell}\circ\mathcal{G}\right)\leq L_\ell\mathfrak{R}_n\left(\mathcal{G}\right)$ according to Talagrand's contraction lemma (Ledoux and Talagrand, 2013). ■

Note that $\ell\left(z\right)-\ell\left(-z\right)=a$ is incorrectly assumed in (Ishida et al., 2017), which causes miscalculation in proof of Lemma 3.

We further describe proof for pairwise classification. Under the assumption that h and $\boldsymbol{\mathcal{L}}_{\mathrm{PC}}$ are equivalent,

$$
\begin{aligned}
&\mathfrak{R}_n\left(\mathcal{H}_{\mathrm{PC}}\right)\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}_{\mathrm{PC}}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h\left(x_i,Y_i\right)\right]\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\widetilde{Y}_i}\widetilde{\mathcal{L}}_{\mathrm{PC}}\left(f(x_i),y\right)\right]\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\widetilde{Y}_i}\sum_{y'\neq y}\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)\right]\\
&=\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)I(y\in\widetilde{Y}_i)\right]\\
&\leq\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,gy'\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)(\alpha_i+1)\right]\\
&\leq\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\left\{\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,gy'\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\alpha_i\sigma_i\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)\right]\right.\\
&\quad\left.+\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,gy'\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)\right]\right\}\\
&\leq\sum_{y\in\mathcal{Y}}\sum_{y'\neq y}\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,g_{y'}\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(g_y(x_i)-g_{y'}(x_i)\right)\right]
\end{aligned}
$$

Let we define $\mathcal{G}_{g_y,g_{y'}}:=\left\{x\mapsto g_y(x)-g_{y'}(x)|g_y,g_{y'}\in\mathcal{G}\right\}$, then:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,g_{y'}\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\ell}\left(g_y(x_i)-g_y(x_i)\right)\right]\\
&=\mathfrak{R}_n\left(\widetilde{\ell}\circ\mathcal{G}_{g_y,g_{y'}}\right)\\
&\leq L_\ell\mathfrak{R}_n\left(\mathcal{G}_{g_y,g_{y'}}\right)\\
&=L_\ell\mathbb{E}_{\mathcal{S}}\mathbb{E}_\sigma\left[\sup_{g_y,g_{y'}\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_y(x_i)-g_{y'}(x_i)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq L_\ell \mathbb{E}_{\mathcal{S}} \mathbb{E}_\sigma \left[\sup_{g_y \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g_y(x_i) \right] + L_\ell \mathbb{E}_{\mathcal{S}} \mathbb{E}_\sigma \left[\sup_{g_{y'} \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (-\sigma_i) g_{y'}(x_i) \right] \\
&= 2 L_\ell \mathfrak{R}_n \left(\mathcal{G} \right)
\end{aligned}
$$

The third equality holds because σ_i and $-\sigma_i$ are drawn from the same probabilistic distribution. Then,

$$
\mathfrak{R}_n \left(\mathcal{H}_{\mathrm{PC}} \right) \leq 2K(K-1) L_\ell \mathfrak{R}_n \left(\mathcal{G} \right)
$$

■

Appendix F. Proof of Theorem 4

We only describe proof for one-versus-all classification; proof for pairwise classification is similar. We substitute the j th data (x_j, Y_j) in $\mathcal{S}$ with any data (x'_j, Y'_j), and define the data set as $\mathcal{S}'$. Let a set of empirical discrimination functions and empirical risk for $\mathcal{S}'$ be $\mathcal{G}' := \{g'\}$ and $\hat{R}'(f)$ respectively. Then the following formulation holds due to Lemma 2.

$$
\begin{aligned}
&\sup_{g_1, \cdots, g_K \in \mathcal{G}} \left(\hat{R}(f) - R(f) \right) - \sup_{g'_1, \cdots, g'_K \in \mathcal{G}'} \left(\hat{R}'(f) - R(f) \right) \\
&\leq \frac{K-1}{n(K-N)} \sup_{g_1, \cdots, g_K \in \mathcal{G}} \inf_{g'_1, \cdots, g'_K \in \mathcal{G}'} \left\{ \left(\boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x_j), Y_j \right) - \boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x'_j), Y'_j \right) \right) \right\} \\
&\leq \frac{K-1}{n(K-N)} \| \boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \|_\infty \\
&\leq \frac{K \widetilde{N}}{n(K-N)}
\end{aligned}
$$

According to McDiarmid's inequality(McDiarmid, 1989), for any integer $\delta > 0$ the following formulation holds with a probability at least $1 - \delta/2$.

$$
\sup_{g_1, \cdots, g_K} \left(\hat{R}(f) - R(f) \right) - \mathbb{E} \left[\sup_{g_1, \cdots, g_K} \left(\hat{R}(f) - R(f) \right) \right] \leq \frac{K \widetilde{N}}{2(K-N)} \sqrt{\frac{2 \ln(2/\delta)}{n}}
$$

Let $\mathcal{S}' := \{(x'_i, Y'_i)\}_{i=1}^n$ be any dataset where each data is drawn from the data-generation probability model $P_N(x, Y)$. Due to $R(f) = \mathbb{E} \left[\hat{R}(f) \right]$,

$$
\begin{aligned}
&\mathbb{E} \left[\sup_{g_1, \cdots, g_K \in \mathcal{G}} \left(\hat{R}(f) - R(f) \right) \right] \\
&= \frac{K-1}{K-N} \mathbb{E}_{\mathcal{S}} \left[\sup_{g_1, \cdots, g_K \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x_i), Y_i \right) - \mathbb{E}_{\mathcal{S}'} \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x'_i), Y'_i \right) \right] \right) \right] \\
&\leq \frac{K-1}{K-N} \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{S}'} \left[\sup_{g_1, \cdots, g_K \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x_i), Y_i \right) - \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\mathcal{L}}_{\mathrm{OVA}} \left(f(x'_i), Y'_i \right) \right) \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{K-1}{K-N}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}\left(f(x_i),Y_i\right)+\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}\left(f(x_i'),Y_i'\right)\right)\right]\\
&\leq \frac{2(K-1)}{K-N}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g_1,\cdots,g_K\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}\left(f(x_i),Y_i\right)\right]\\
&= \frac{2(K-1)}{K-N}\mathfrak{R}_n\left(\mathcal{H}_{\mathrm{OVA}}\right)\\
&\leq \frac{2K(K+\widetilde{N})}{K-N}L_{\ell}\mathfrak{R}_n\left(\mathcal{G}\right)
\end{aligned}
$$

The second equality holds because $\boldsymbol{\mathcal{L}}\left(f(x_i,Y_i)\right)$ and $\sigma_i\boldsymbol{\mathcal{L}}\left(f(x_i,Y_i)\right)$ are drawn from the same probabilistic distribution; similar for $\boldsymbol{\mathcal{L}}\left(f(x_i',Y_i')\right)$. The last inequality holds due to Lemma 3. $\hat{R}(\hat{f})\leq\hat{R}(f^*)$ holds according to (20), thus,

$$
\begin{aligned}
\mathcal{E}_N &= \left(\hat{R}(\hat{f})-\hat{R}(f^*)\right)+\left(R(\hat{f})-\hat{R}(\hat{f})\right)+\left(\hat{R}(f^*)-R(f^*)\right)\\
&\leq 2\sup_{g_1,\cdots,g_K\in\mathcal{G}}\left|\hat{R}(f)-R(f)\right|\\
&\leq \frac{4K(K+\widetilde{N})}{K-N}L_{\ell}\mathfrak{R}_n\left(\mathcal{G}\right)+\frac{K\widetilde{N}}{K-N}\sqrt{\frac{2\ln(2/\delta)}{n}}
\end{aligned}
$$

■

Appendix G. Comparison with Existing Work

Our major contribution is the finding that loss functions reflecting the properties of label space naturally bridge ordinary-label learning and complementary-label learning. However, unlike (Ishida et al., 2019; Feng et al., 2020), it is also true that satisfaction of additivity and duality is a strict limitation. In this section we present an experiment with a loss function which does not satisfy the assumption to evaluate the limitation of our work.

Here we introduce a loss function $\boldsymbol{\mathcal{L}}_{\mathrm{CE}}$ as follows:

$$
\boldsymbol{\mathcal{L}}_{\mathrm{CE}}\left(f(x),Y\right)=-\sum_{y\in Y}\log\left(\frac{\exp(g_y(x))}{\sum_{y'\in\mathcal{Y}}\exp(g_{y'}(x))}\right)
$$

$\boldsymbol{\mathcal{L}}_{\mathrm{CE}}$ is represented as an additive form of cross entropy loss, the most commonly used loss function. Unlike $\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}$ or $\boldsymbol{\mathcal{L}}_{\mathrm{PC}}$, learning with $\boldsymbol{\mathcal{L}}_{\mathrm{CE}}$ is outside the scope of our framework because additivity and duality are not satisfied.

We performed an experiment with $\boldsymbol{\mathcal{L}}_{\mathrm{CE}}$ on the same dataset described in Section 6.1 to evaluate the classification accuracy and error. We set the batch size as 64 and the number of epochs as 300. MLP was adopted for MNIST, Fashion-MNIST, and Kuzushiji-MNIST datasets; *Adam* was used for optimization with weight decay of 10^{-4} and learning rate of 5×10^{-5}. DenseNet was adopted for CIFAR-10 dataset; *stochastic gradient descent* was used for optimization with weight decay of 5×10^{-4} and initial learning rate of 5×10^{-2}

Table 4: Experimental classification accuracies for 10 and 5 class classification (%). The experiments were performed 5 times for each case; the mean accuracy and standard deviation are presented by the upper and lower values, respectively. The highest accuracy is boldfaced.

	$K = 10$									$K = 5$			
N	1	2	3	4	5	6	7	8	9	1	2	3	4
MNIST	**94.69**	93.70	93.1	92.17	91.11	89.60	86.42	80.99	71.33	**98.05**	97.05	96.02	94.28
(CE)	(±0.13)	(±0.19)	(±0.25)	(±0.50)	(±0.35)	(±0.73)	(±0.79)	(±0.81)	(±0.78)	(±0.10)	(±0.14)	(±0.52)	(±0.49)
Fashion	**86.81**	85.34	84.97	83.75	82.93	80.60	77.89	73.83	63.05	**87.21**	85.91	84.16	81.63
(CE)	(±0.43)	(±0.23)	(±0.36)	(±0.36)	(±0.30)	(±0.66)	(±0.35)	(±0.63)	(±1.91)	(±0.27)	(±0.65)	(±0.45)	(±0.66)
Kuzushiji	**77.19**	71.38	69.22	66.43	63.05	57.31	51.62	42.25	30.99	**82.01**	78.23	73.84	65.58
(CE)	(±0.53)	(±0.60)	(±0.53)	(±0.76)	(±0.36)	(±0.36)	(±1.13)	(±1.22)	(±0.51)	(±0.48)	(±0.77)	(±1.12)	(±2.91)
CIFAR-10	**73.23**	56.84	43.00	36.03	32.57	35.47	40.18	40.88	29.70	**74.81**	64.69	58.54	49.43
(CE)	(±0.20)	(±0.52)	(±0.55)	(±0.77)	(±1.25)	(±1.76)	(±0.45)	(±0.89)	(±0.53)	(±0.36)	(±0.93)	(±0.80)	(±1.97)

Table 5: Experimental classification errors for 10 and 5 class classification ($\times 10^{-3}$). The experiments were performed 5 times for each case; the mean error and standard deviation are presented by the upper and lower values, respectively. The highest error is boldfaced.

	$K = 10$									$K = 5$			
N	1	2	3	4	5	6	7	8	9	1	2	3	4
MNIST	0.07	0.40	0.49	0.49	0.61	0.68	0.51	0.46	**0.86**	0.02	0.43	**0.51**	0.33
(CE)	(±0.01)	(±0.16)	(±0.22)	(±0.23)	(±0.18)	(±0.11)	(±0.27)	(±0.23)	(±0.26)	(±0.02)	(±0.12)	(±0.15)	(±0.20)
Fashion	0.08	0.39	0.39	0.41	0.34	0.52	0.42	0.38	**0.75**	0.05	0.22	0.30	**0.48**
(CE)	(±0.06)	(±0.22)	(±0.13)	(±0.23)	(±0.19)	(±0.22)	(±0.16)	(±0.29)	(±0.49)	(±0.02)	(±0.14)	(±0.04)	(±0.24)
Kuzushiji	0.33	0.36	0.27	0.23	0.28	0.28	0.23	0.65	**0.92**	0.18	0.26	0.27	**0.46**
(CE)	(±0.13)	(±0.07)	(±0.14)	(±0.14)	(±0.13)	(±0.09)	(±0.10)	(±0.30)	(±0.36)	(±0.12)	(±0.15)	(±0.16)	(±0.30)
CIFAR-10	1.31	1.34	1.49	1.55	**1.72**	1.01	0.28	0.08	0.07	0.04	**0.25**	0.22	0.13
(CE)	(±0.21)	(±0.17)	(±0.15)	(±0.14)	(±0.08)	(±0.10)	(±0.08)	(±0.03)	(±0.02)	(±0.03)	(±0.07)	(±0.06)	(±0.02)

which was halved every 30 epochs. The results of accuracy and error are listed in Table 4 and 5 correspondingly.

Compared with the results in Table 2 and 3, the classifier with $\boldsymbol{\mathcal{L}}_{\mathrm{CE}}$ tends to perform better than those of $\boldsymbol{\mathcal{L}}_{\mathrm{OVA}}$ and $\boldsymbol{\mathcal{L}}_{\mathrm{PC}}$. This shows that although the assumption of additivity and duality naturally reflects the relationship between ordinary-label and complementary-label, it could be a limitation on performance in a context of practical use.