A Experiment Details

A.1 Datasets

We perform experiments and compare against the KL loss of Dulac-Arnold et al. [9] and LLPVAT of Tsai and Lin [46] on three benchmark datasets of image classification: the “letter" split of EMNIST [6], SVHN [29], and CIFAR10 [17]. We also compare our methods against LLPGAN of Liu et al. [22] on SVHN and CIFAR10. To generate each bag, we first sample a label proportion γ from the uniform distribution on Δ^C and then sample data points without replacement using a multinomial distribution with parameter γ. The generated bags have fixed and equal sizes in $\{32, 64, 128, 256, 512, 1024, 2048\}$. For SVHN and CIFAR10, $32 \times 1280 = 40960$ data points are sampled for every bag size. For EMNIST, the number of sampled data points is $32 \times 3328 = 106496$.

A.2 Architecture and Hyperparameters

To compare our methods against the KL loss and LLPVAT, we train Wide ResNet-16-4 [52], ResNet18 [11], and VGG16 [42] with the hyperparameters suggested in the original papers. For the comparison against LLPGAN, we use the discriminator architecture proposed in Liu et al. [22] and the hyperparameters suggested in their code[3].It should be noted that KL, LLPVAT, and LLPGAN are all required to backpropagate on minibatches of bags and our method does not have such constraint. For all methods, to avoid overfitting, we apply a standard data augmentation procedure: 4 pixels with value 0 are padded on each side, and a crop of the original size is randomly sampled from the padded image or its horizontal flip.

A.2.1 Wide ResNet-16-4

For all datasets, we use SGD with Nesterov momentum with weight decay set to 0.0005, dampening to 0, and momentum to 0.9. The minibatch size is set to 128 for our method. On CIFAR, the initial learning rate is set to 0.01, which is divided by 5 at 60, 120 and 160 epochs, and the network is trained for total 200 epochs. On SVHN and EMNIST, the initial learning rate is set to 0.01, which is divided by 10 at 80 and 120 epochs, and the network is trained for 160 epochs. The dropout probability is 0.3 for CIFAR and 0.4 for both SVHN and EMNIST.

A.2.2 ResNet18

We use SGD and weight decay is set to 0.0001 and momentum to 0.9. The minibatch size is set to 128 for our method. The model is trained for 500 epochs for all datasets. The learning rate is initialized to be 0.1 and divided by 10 at 250 and 375 epochs.

A.2.3 VGG16

We use SGD and weight decay is set to 0.0005 and momentum to 0.9. The minibatch size is set to 256 for our method. Dropout ratio is set to 0.5 for the first two fully-connected layers. The learning rate was initially set to 0.01. We train the model for 74 epochs in total. In the original paper of VGG16 [42], the learning rate is decreased when validation accuracy stops improving and it is decreased 3 times in total. In our experiment, while we do not assume access to fully labeled validation dataset, we divide the learning rate by 10 at 19, 37, and 56 epochs.

A.2.4 LLPGAN's discriminator

The neural network is trained for 3000 epochs and optimized by Adam [15] with a learning rate 0.0003. The minibatch size is set to 128 for our method. The β_1 and β_2 parameters for Adam are set to be 0.5 and 0.999, respectively.

A.3 KL Loss

Recall $C \in \mathbb{N}$ denotes the number of classes and $\mathcal{X}$ denotes the feature space. Let f be a function that maps $\mathcal{X}$ to Δ^C, and $M \in \mathbb{N}$ be the total number of bags. Let $\{(b_i, \hat{\gamma}_i)\}_{i=1}^M$ be

---

[3] https://github.com/liujiabin008/LLP-GAN

the bags and empirical label proportions where $b_i = (X_1^i, X_2^i, \ldots, X_{m_i}^i)$ and m_i is the size of bag b_i. The KL loss of Dulac-Arnold et al. [9] seeks to minimize the empirical objective $\bar{\mathcal{L}}_{prop} = -\frac{1}{CM}\sum_{i=1}^{M}\sum_{c=1}^{C}\hat{\gamma}_i(c)\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i} f_c(X_j^i)\right)$ over the function f in some space $\mathcal{F}_0$. The $\hat{\gamma}_i(c)$ is the label proportion of c-th entry in bag i and f_c is the c-th entry of the output of f. In practice, when $\mathcal{F}_0$ is softmax composed with neural networks of certain architecture, the objective is optimized by stochastic gradient descent (SGD) with “minibatches of bags". For a minibatch of size B, B bags $b_{i_1}, \ldots, b_{i_B}$ are sampled and SGD backpropagates the gradients of $\mathcal{L}_{prop}(b_{i_1}, \ldots, b_{i_B}) = -\frac{1}{CB}\sum_{k=1}^{B}\sum_{c=1}^{C}\hat{\gamma}_{i_k}(c)\log\left(\frac{1}{m_{i_k}}\sum_{j=1}^{m_{i_k}} f_c(X_j^{i_k})\right)$. In our experiments, we follow the code of Tsai and Lin [46][4] and set $B = 2$ (Dulac-Arnold et al. [9] also use minibatches of bags but do not specify B). While optimizing neural networks with the KL loss on GPU nodes, the gradients of all data points in the minibatch of bags need to be stored in the GPU memory simultaneously. So KL loss can potentially exceed GPU memory when bag size increases. In this situation, we report *Out of RAM* in the tables.

A.4 LLPVAT

Let f and $\mathcal{L}_{prop}$ be defined as in A.3. Let D_{KL} denote the KL divergence. The LLPVAT algorithm of Tsai and Lin [46] computes the perturbed examples $\hat{x} = x + r_{adv}$ where

$$r_{adv} = \underset{r:\|r\|_2 \le \epsilon}{\arg\max}\, D_{KL}(f(x) \,||\, f(x + r_{adv})).$$

Given a minibatch of B bags $b_{i_1}, \ldots, b_{i_B}$, their consistency loss is defined to be

$$\mathcal{L}_{cons}(b_{i_1}, \ldots, b_{i_B}) = \sum_{k=1}^{B}\frac{1}{|b_{i_k}|}\sum_{x \in b_{i_k}} D_{KL}(f(x) \,||\, f(x + r_{adv})).$$

For each minibatch in the t-th epoch, the LLPVAT algorithm updates the parameters of neural networks with the gradients of the loss $\mathcal{L}(b_{i_1}, \ldots, b_{i_B}) = \mathcal{L}_{prop}(b_{i_1}, \ldots, b_{i_B}) + w(t)\mathcal{L}_{cons}(b_{i_1}, \ldots, b_{i_B})$ where $w(t)$ is a ramp-up function for increasing the weight of consistency regularization. Following the LLPVAT paper, we set ϵ to 1 for both SVHN and EMNIST and set ϵ to 6 for CIFAR10. We follow the code of Tsai and Lin [46][5] to implement $w(t)$ and set minibatch size B to be 2. Like the KL loss, LLPVAT can potentially exceed GPU memory. In this situation, we report *Out of RAM* in the tables.

A.5 LLPGAN

Let bags b_i and $\bar{\mathcal{L}}_{prop}$ be defined as in subsection A.3. The LLPGAN model of Liu et al. [22] consists of a generator g and a discriminator f. The discriminator f is a convolutional neural network and we denote its convolutional layers as f_{conv}. The generator g maps a random noise to the image space and the discriminator maps an image to Δ^{C+1} where the fake images output by the generator are supposed to be classified as the $(C+1)$-th class. Let n be the total number of feature vectors in all bags, *i.e.*, $n = \sum_{i=1}^{M} m_i$ and $\left\{Z_i^j, i \in \mathbb{N}_M, j \in \mathbb{N}_{m_i}\right\}$ be random noise vectors sampled from a fixed distribution. The discriminator loss is defined as

$$\bar{\mathcal{L}}_D = -\sum_{i=1}^{M}\sum_{j=1}^{m_i}\frac{1}{m_i}\log\left(\sum_{c \le C} f_c(X_j^i)\right) - \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{m_i}\log\bigl(f_{C+1}(g(Z_j^i))\bigr)$$
$$-\frac{1}{MC}\sum_{i=1}^{M}\sum_{c=1}^{C}\sum_{j=1}^{m_i}\frac{\hat{\gamma}_i(c)}{m_i}\log\bigl(f_c\bigl(X_j^i\bigr)\bigr),$$

where the last term $-\frac{1}{MC}\sum_{i=1}^{M}\sum_{c=1}^{C}\sum_{j=1}^{m_i}\frac{\hat{\gamma}_i(c)}{m_i}\log\bigl(f_c\bigl(X_j^i\bigr)\bigr)$ is proposed as an upper bound of $\bar{\mathcal{L}}_{prop}$. The generator loss is defined as

$$\bar{\mathcal{L}}_G = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{m_i}\bigl\|f_{conv}(X_j^i) - f_{conv}(g(Z_j^i))\bigr\|_2^2.$$

---

[4] https://github.com/kevinorjohn/LLP-VAT

[5] See footnote 4

Given a minibatch of B bags $b_{i_1}, \ldots, b_{i_B}$, the minibatch version of the discriminator loss is

$$\begin{aligned}\mathcal{L}_D(b_{i_1}, \ldots, b_{i_B}) = &- \sum_{k=1}^{B} \sum_{j=1}^{m_{i_k}} \frac{1}{m_{i_k}} \log \left(\sum_{c \leq C} f_c(X_j^{i_k}) \right) \\ &- \frac{1}{\sum_{k=1}^{B} m_{i_k}} \sum_{i=1}^{B} \sum_{j=1}^{m_{i_k}} \log\big(f_{C+1}(g(Z_j^{i_k}))\big) \\ &- \frac{1}{BC} \sum_{k=1}^{B} \sum_{c=1}^{C} \sum_{j=1}^{m_{i_k}} \frac{\hat{\gamma}_i(c)}{m_{i_k}} \log\big(f_c(X_j^{i_k})\big).\end{aligned}$$

The minibatch version of generator loss is

$$\mathcal{L}_G(b_{i_1}, \ldots, b_{i_B}) = \frac{1}{\sum_{k=1}^{B} m_{i_k}} \sum_{i=1}^{B} \sum_{j=1}^{m_{i_k}} \left\| f_{conv}(X_j^{i_k}) - f_{conv}(g(Z_j^{i_k})) \right\|_2^2.$$

So the training process of LLPGAN can be described as follows: in each epoch, for a minibatch $(b_{i_1}, \ldots, b_{i_B})$,

1. Sample random noise $Z_j^{i_k}$ for $k \in \mathbb{N}_B$ and $j \in \mathbb{N}_{m_{i_k}}$.
2. Fix g and perform gradient descent on parameters of f in $\mathcal{L}_D$.
3. Fix f and perform gradient descent on parameters of g in $\mathcal{L}_G$.

The code [6] of the original paper implements f as a neural network which outputs a vector in $\mathbb{R}^C$ followed by a $(C+1)$-way softmax with the $(C+1)$-th input fixed to be 0. We follow this practice in our implementation of LLPGAN. We also follow the code of original paper and set B to be 1. While LLPGAN could potentially exceed GPU memory while bag size increases as well, this did not happen in our experiments.

A.6 Implementation Details of LLPFC

The performance of LLPFC algorithms benefit from re-partitioning of bags periodically. We randomly repartition the bags into groups every 20 epochs for WideResNet-16-4, ResNet-18, and LLPGAN discriminator and every 5 epochs for VGG16.

A.7 Experiments in Binary Setting

We carry out an extra set of experiments with kernel methods on binary classification tasks, comparing against InvCal [36], alter-$\propto$SVM [50], and LMMCM [39]. We run our experiments on the exact same datasets used by Scott and Zhang [39] and directly compare against the results presented in their paper. We implement LLPFC-uniform and LLPFC-approx with rbf kernel models and logistic loss by modifying the code provided by Scott and Zhang [39] at `https://github.com/Z-Jianxin/Learning-from-Label-Proportions-A-Mutual-Contamination-Framework`. We run experiments in the same settings of Scott and Zhang [39]. The model is solved by L-BFGS. We compute the kernel parameter by $\frac{1}{d * Var(X)}$ where d is the number of features and $Var(X)$ is the variance of the data matrix. The regularization parameter $\lambda \in \{1, 10^{-1}, 10^{-2}, \ldots, 10^{-5}\}$ is chosen by 5-fold cross validation, using the empirical risk provided in Algorithm 2 and Algorithm 3, respectively. We evaluate the area under the ROC curve (AUC) and report the results in table 5. We bold the largest mean AUC for that experimental setting. Each of LLPFC-uniform, LLPFC-approx, and LMMCM achieves the highest AUC among all the methods in 5 settings. LLPFC-uniform also beats the three competitors from Scott and Zhang [39] in 10 out 16 settings.

---

[6] See footnote 3

Table 5: AUC. Column header indicates bag size.

Data set, LP dist	Method	8	32	128	512
Adult, $[0, \frac{1}{2}]$	InvCal	0.8720 ± 0.0035	0.8672 ± 0.0067	0.8537 ± 0.0101	0.7256 ± 0.0159
	alter-$\propto$SVM	0.8586 ± 0.0185	0.7394 ± 0.0686	0.7260 ± 0.0953	0.6876 ± 0.1219
	LMMCM	0.8728 ± 0.0019	**0.8693** ± **0.0047**	**0.8669** ± **0.0041**	**0.8674** ± **0.0040**
	LLPFC-uniform	**0.8751** ± **0.0022**	0.8627 ± 0.0034	0.8616 ± 0.0057	0.8594 ± 0.0047
	LLPFC-approx	0.8676 ± 0.0042	0.8540 ± 0.0052	0.8509 ± 0.0094	0.8478 ± 0.0096
Adult, $[\frac{1}{2}, 1]$	InvCal	0.8680 ± 0.0021	0.8598 ± 0.0073	0.8284 ± 0.0093	0.7480 ± 0.0500
	alter-$\propto$SVM	0.8587 ± 0.0097	0.7429 ± 0.1473	0.8204 ± 0.0318	0.7602 ± 0.1215
	LMMCM	0.8584 ± 0.0164	0.8644 ± 0.0052	0.8601 ± 0.0045	0.8500 ± 0.0186
	LLPFC-uniform	0.8693 ± 0.0036	**0.8666** ± **0.0047**	**0.8636** ± **0.0040**	**0.8587** ± **0.0136**
	LLPFC-approx	**0.8723** ± **0.0014**	0.8630 ± 0.0069	0.8560 ± 0.0103	0.8538 ± 0.0193
MAGIC, $[0, \frac{1}{2}]$	InvCal	**0.8918** ± **0.0076**	0.8574 ± 0.0079	0.8295 ± 0.0139	0.8133 ± 0.0109
	alter-$\propto$SVM	0.8701 ± 0.0026	0.7704 ± 0.0818	0.7753 ± 0.0207	0.6851 ± 0.1580
	LMMCM	0.8909 ± 0.0077	**0.8799** ± **0.0113**	**0.8753** ± **0.0157**	0.8734 ± 0.0092
	LLPFC-uniform	0.8575 ± 0.0644	0.8751 ± 0.0158	0.8715 ± 0.0066	**0.8761** ± **0.0157**
	LLPFC-approx	0.8829 ± 0.0135	0.8590 ± 0.0256	0.8721 ± 0.0054	0.8711 ± 0.0155
MAGIC, $[\frac{1}{2}, 1]$	InvCal	0.8936 ± 0.0066	0.8612 ± 0.0056	0.8180 ± 0.0092	0.8215 ± 0.0136
	alter-$\propto$SVM	0.8689 ± 0.0135	0.8219 ± 0.0218	0.8179 ± 0.0487	0.7949 ± 0.0478
	LMMCM	0.8911 ± 0.0083	0.8790 ± 0.0091	0.8684 ± 0.0046	0.8567 ± 0.0292
	LLPFC-uniform	0.8985 ± 0.0054	0.8851 ± 0.0113	0.8844 ± 0.0101	0.8765 ± 0.0113
	LLPFC-approx	**0.9011** ± **0.0034**	**0.8990** ± **0.0122**	**0.8882** ± **0.0088**	**0.8800** ± **0.0114**

B Proofs of Results from Section 3

B.1 Proof of Theorem 5

To prove Theorem 5, we employ the calibration framework of Steinwart [43]. The first lemma of this section establishes an instance of what Steinwart [43] refers to as *uniform calibration*, but in the LLN setting.

Lemma 15. *Let ℓ be a continuous strictly proper loss and T be an invertible column-stochastic matrix. Let L be the* 0-1 *loss. Then* $\forall \epsilon \in \mathbb{R}_+, \exists \delta \in \mathbb{R}_+, s.t. \forall x \in \mathcal{X}, \forall q \in \Delta^C$,

$$\mathcal{C}_{\ell_T, P_T, x}(q) < \mathcal{C}^*_{\ell_T, P_T, x} + \delta \implies \mathcal{C}_{L,P,x}(q) < \mathcal{C}^*_{L,P,x} + \epsilon.$$

Proof of Lemma 15. Write

$$\mathcal{C}_{1,x}(q) := \mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x}$$
$$\mathcal{C}_{2,x}(q) := \mathcal{C}_{\ell_T,P_T,x}(q) - \mathcal{C}^*_{\ell_T,P_T,x}.$$

Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^C$, k and $j \in \{1, \dots, C\}$ be such that $q_k = \max_i q_i$, and $\eta_j(x) = \max_i \eta_i(x)$. Then we have

$$\begin{aligned}
\mathcal{C}_{1,x}(q) &= (1 - \eta_k(x)) - (1 - \eta_j(x)) \\
&= \eta_j(x) - \eta_k(x) \\
&= (\eta_j(x) - q_k) + (q_k - \eta_k(x)) \\
&= \|\eta(x)\|_\infty - \|q\|_\infty + (q_k - \eta_k(x)) \\
&\leq \|\eta(x) - q\|_\infty + \|\eta(x) - q\|_1 \\
&\leq C_{\|\cdot\|} \|\eta(x) - q\|,
\end{aligned}$$

where $C_{\|\cdot\|} > 0$ is a constant depending on the norm $\|\cdot\|$ and the last inequality is implied by the equivalence of norms in finite dimensional space.
Hence,

$$\|\eta(x) - q\| < \frac{\epsilon}{C_{\|\cdot\|}} \implies \mathcal{C}_{1,x}(q) < \epsilon.$$

So it suffices to prove

$$\forall \delta_0 \in \mathbb{R}_+, \exists \delta \in \mathbb{R}_+ s.t. \forall x \in X, \forall q \in \Delta^C, \mathcal{C}_{2,x}(q) < \delta \implies \|q - \eta(x)\| < \delta_0.$$

To this end, assume its negation,

$$\exists \delta_0 \in \mathbb{R}_+, \forall \delta \in \mathbb{R}_+ s.t. \exists x \in X, \exists q \in \Delta^C, \mathcal{C}_{2,x}(q) < \delta \text{ and } \|q - \eta(x)\| \geq \delta_0.$$

Let $\delta_n = \frac{1}{n}$ for each $n \in \mathbb{N}$. We can obtain a sequence $\{(q_i, \eta(x_i))\}_{i=1}^{\infty} \subset K := \{(q,\eta) \in \Delta^C \times \Delta^C : \|q - \eta\| \geq \delta_0\}$ such that $\mathcal{C}_{2,x_i}(q_i) < \delta_i$ for all $i \in \mathbb{N}$. As K is compact, we can extract a convergent subsequence $\{(q_{i_k}, \eta(x_{i_k}))\}_{k=1}^{\infty}$. Let

$$\lim_k (q_{i_k}, \eta(x_{i_k})) = (q^*, \eta^*) \in K.$$

Write

$$\mathcal{C}_2(q,\eta) := \sum_{c=1}^{C} (T\eta)_c (\ell(Tq, c) - \ell(T\eta, c)),$$

so $\mathcal{C}_2$ is continuous and $\mathcal{C}_{2,x}(q) = \mathcal{C}_2(q, \eta(x))$. Therefore,

$$0 = \lim_k \mathcal{C}_{2,x_{i_k}}(q_{i_k}) = \lim_k \mathcal{C}_2(q_{i_k}, \eta(x_{i_k})) = \mathcal{C}_2(q^*, \eta^*),$$

which contradicts the strict properness of ℓ since $q^* \neq \eta^*$. □

This lemma establishes what may be viewed as a pointwise notion of consistency: For each fixed $x \in \mathcal{X}$, the target excess 0/1-inner risk (defined $w.r.t.$ P) can be made arbitrarily small by making the surrogate excess ℓ_T-inner risk (defined $w.r.t.$ P_T) sufficiently small.

Now let $\epsilon \in [0,\infty]$, and denote $A(\epsilon) :=$

$$\{\delta \in [0,\infty] : \forall x \in \mathcal{X}, \forall q \in \Delta^C, \mathcal{C}_{\ell_T,P_T,x}(q) < \mathcal{C}^*_{\ell_T,P_T,x} + \delta \implies \mathcal{C}_{L,P,x}(q) < \mathcal{C}^*_{L,P,x} + \epsilon\}.$$

Define the function

$$\delta : [0,\infty] \to [0,\infty], \;\; \delta(\epsilon) = \sup_{\delta \in A(\epsilon)} \delta.$$

The following properties immediately follow from the definition and Lemma 15:

- $\delta(0) = 0$ and $\delta(\epsilon) > 0$ if $\epsilon > 0$.
- $\delta(\cdot)$ is monotone non-decreasing.
- $A(\epsilon) = [0, \delta(\epsilon)]$

The function δ is a reasonable candidate for the sought after function θ, but it is not necessarily invertible since it might not be strictly increasing. To address this we introduce the following.

Definition 16. *Let $I \subset \mathbb{R}$ be an interval and let $g : I \to [0,\infty]$ be a function. Then the* Fenchel-Legendre biconjugate $g^{**} : I \to [0,\infty]$ *of g is the largest convex function $h : I \to [0,\infty]$ satisfying $h \leq g$.*

We are now prepared to prove theorem 17 of which theorem 5 is a direct corollary.

Theorem 17. *Let ℓ be a continuous strictly proper loss and T be an invertible column-stochastic matrix. Let L be the* 0-1 *loss and let $\delta(\cdot)$ be the function defined above. Assume $\mathcal{R}^*_{\ell_T,P_T} < \infty$. Then for all P,*

$$\forall f \in \mathcal{F}, \delta^{**}_{|[0,1]}(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}) \leq \mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T}$$

where $\delta^{**}_{|[0,1]}$ denotes the Fenchel-Legendre biconjugate of the restriction $\delta_{|[0,1]}$.

Proof of Theorem 17. Write

$$\mathcal{C}_{1,x}(q) := \mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x}$$
$$\mathcal{C}_{2,x}(q) := \mathcal{C}_{\ell_T,P_T,x}(q) - \mathcal{C}^*_{\ell_T,P_T,x}.$$

Then,

$$\forall x \in \mathcal{X}, \forall p,q \in \Delta^C, \mathcal{C}_{2,x}(p) < \delta(\mathcal{C}_{1,x}(q)) \implies \mathcal{C}_{1,x}(p) < \mathcal{C}_{1,x}(q).$$

By letting $p = q$, we have $\forall x \in \mathcal{X}, \forall q \in \Delta^C, \mathcal{C}_{2,x}(q) \geq \delta(\mathcal{C}_{1,x}(q))$. Fix $f \in \mathcal{F}$,

$$\begin{aligned}
\delta^{**}_{|[0,1]}(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}) &= \delta^{**}_{|[0,1]}\left(\int_{\mathcal{X}} \mathcal{C}_{1,x}(f(x)) dP_X(x)\right) && (2)\\
&\leq \int_{\mathcal{X}} \delta^{**}_{|[0,1]}(\mathcal{C}_{1,x}(f(x))) dP_X(x) && (3)\\
&\leq \int_{\mathcal{X}} \mathcal{C}_{2,x}(f(x)) dP_X(x) && (4)\\
&= \mathcal{R}_{\ell_T,P_T}(f) - \mathcal{R}^*_{\ell_T,P_T} && (5)
\end{aligned}$$

(2) follows the fact $\mathcal{R}_{L,P}(f) = \int_{\mathcal{X}} \mathcal{C}_{L,P,x}(f(x))dP_X(x)$ and $\mathcal{R}^*_{L,P} = \int_{\mathcal{X}} \mathcal{C}_{L,P,x}(\eta(x))dP_X(x) = \int_{\mathcal{X}} \mathcal{C}^*_{L,P,x}dP_X(x)$. (3) is implied by Jensen's inequality and the convexity of $\delta^{**}_{|[0,1]}$ and (4) by the fact $\delta^{**}_{|[0,1]}(\cdot) \leq \delta(\cdot)$. (5) follows $\mathcal{R}_{\ell_T,P_T}(f) = \int_{\mathcal{X}} \mathcal{C}_{\ell_T,P_T,x}(f(x))dP_X(x)$ and $\mathcal{R}^*_{\ell_T,P_T} = \int_{\mathcal{X}} \mathcal{C}_{\ell_T,P_T,x}(\eta(x))dP_X(x) = \int_{\mathcal{X}} \mathcal{C}^*_{\ell_T,P_T,x}dP_X(x) < \infty$. ☐

B.2 Proof of Proposition 6

For the reader's convenience, we restate Proposition 6 below:

Proposition 18. *Let $T \in \mathbb{R}^{C\times C}$ be an invertible, column-stochastic matrix. Define $\underline{\theta}_T : [0,\infty] \to [0,\infty]$ by*

$$\underline{\theta}_T(\epsilon) = \frac{1}{2}\frac{\epsilon^2}{\|T^{-1}\|_1^2}.$$

Then for all $x \in X$ and $q \in \Delta^C$, we have

$$\underline{\theta}_T(\mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x}) \leq \mathcal{C}_{\ell_T,P_T,x}(q) - \mathcal{C}^*_{\ell_T,P_T,x}.$$

Below, let ℓ denote the *log loss* $\ell^{log}(q,c) = -\log q_c$ and L denote the $0-1$ loss:

$$L : \Delta^C \times Y \to \{0,1\}, \qquad L(q,c) = \mathbb{1}_{\{c \neq \min\{\arg\max q\}\}}.$$

To proceed with the proof, we first introduce some notations and useful results. For $p, q \in \Delta^C$, define

$$\underline{\ell}(q,p) := \mathbb{E}_{y\sim p}\ell(q,y) = \sum_{i=1}^{C} -p_i \log q_i. \tag{6}$$

The above quantity is often referred to as the *cross entropy* of q relative to p. Next, since the log loss is proper [49], we have

$$\inf_{q\in\Delta^C} \underline{\ell}(q,p) = \underline{\ell}(p,p). \tag{7}$$

The *Kullback-Leibler (KL) divergence* between $p, q \in \Delta^c$ is defined as

$$\mathtt{KL}(p\|q) = \underline{\ell}(p,q) - \underline{\ell}(p,p). \tag{8}$$

In the literature, the KL divergence is often presented as $\mathtt{KL}(p\|q) = \sum_{i=1}^C p_i \log\left(\frac{p_i}{q_i}\right)$ which is easily shown to be equivalent to (8). We now rewrite the right-hand side of the inequality in Proposition 18 in terms of the KL divergence:

Lemma 19. *Let $p := P(\cdot|x) \in \Delta^C$. Then*

$$\mathcal{C}_{\ell_T,P_T,x}(q) - \mathcal{C}^*_{\ell_T,P_T,x} = \mathtt{KL}(Tp\|Tq). \tag{9}$$

Proof of Lemma 19. By definition, we have $Tp = P_T(\cdot|x)$. Unwinding the definitions, we have

$$\mathcal{C}_{\ell_T,P_T,x}(q) = \mathbb{E}_{y\sim P_T(\cdot|x)}\ell_T(q,y) = \mathbb{E}_{y\sim Tp}\ell(Tq,y) = \underline{\ell}(Tq,Tp).$$

Furthermore,

$$\mathcal{C}^*_{\ell_T,P_T,x} = \inf_{q\in\Delta^C} \mathbb{E}_{y\sim P_T(\cdot|x)}\ell_T(q,y) = \inf_{q\in\Delta^C} \underline{\ell}(Tq,Tp) = \underline{\ell}(Tp,Tp)$$

where the last equality follows from (7). Now, (9) follows immediately from (8). ☐

Next, we focus on the term $\mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x}$ on the left-hand side in Proposition 18. Analogous to (6), we define

$$\underline{L}(q,p) := \mathbb{E}_{y\sim p}L(q,y) = \sum_{c=1}^{C} p_c \mathbb{1}_{\{c\neq\min\{\arg\max q\}\}} = 1 - p_{\min\{\arg\max q\}}. \tag{10}$$

The $0-1$ loss is also proper and

$$\inf_{q\in\Delta^C} \underline{L}(q,p) = \underline{L}(p,p) = 1 - p_{\min\{\arg\max p\}} = 1 - \max_c p_c. \tag{11}$$

Unwinding the definition, we have

$$\mathcal{C}_{L,P,x}(q) = \mathbb{E}_{y \sim P(\cdot|x)} L(q,y) = \underline{L}(q,p)$$

and

$$\mathcal{C}^*_{L,P,x} = \inf_{q \in \Delta^C} \mathbb{E}_{y \sim P(\cdot|x)} L(q,y) = \inf_{q \in \Delta^C} \underline{L}(q,p) = \underline{L}(p,p).$$

Thus,

$$\mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x} = \underline{L}(q,p) - \underline{L}(p,p). \tag{12}$$

Thus, by (9) and (12), we only need to focus on comparing $\underline{L}(q,p) - \underline{L}(p,p)$ with $\mathtt{KL}(Tp\|Tq)$. This is facilitated by the 1-norm $\|\cdot\|_1$ and the next two results. The first is by Pinsker [31]:

Theorem 20 (Pinsker inequality)**.** *Let $\|\cdot\|_1$ be the 1-norm on $\mathbb{R}^C$. Then for all $p, q \in \Delta^C$, we have*

$$\mathtt{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2.$$

The second one is widely-known in the literature. For the sake of completeness, we provide a proof using our notations:

Lemma 21. *Let $p, q \in \Delta^C$ be arbitrary. Then $\|p-q\|_1 \geq \underline{L}(q,p) - \underline{L}(p,p)$.*

Proof of Lemma 21. Let $i := \min\{\arg\max p\}$ and $j := \min\{\arg\max q\}$. Then by (10) and (11), we have

$$\underline{L}(q,p) - \underline{L}(p,p) = 1 - p_j - (1 - p_i) = p_i - p_j.$$

On the other hand, note that

$$\begin{aligned}
\|p-q\|_1 &= \sum_{i=1}^{C} |p_i - q_i| \\
&\geq |p_i - q_i| + |q_j - p_j| \\
&\geq |p_i - p_j + q_j - q_i| && \because \text{triangle inequality} \\
&= p_i - p_j + q_j - q_i && \because p_i - p_j \geq 0 \text{ and } q_j - q_i \geq 0 \\
&\geq p_i - p_j
\end{aligned}$$

as desired. □

Finally, we need one more result to take into account the presence of the stochastic matrix T when applying Pinsker inequality to lower bound $\mathtt{KL}(Tp\|Tq)$:

Lemma 22. *Let $M \in \mathbb{R}^{C \times C}$ be a matrix and let $\|\cdot\|$ be a norm on $\mathbb{R}^C$. Suppose that M is non-singular. Then*

$$\inf_{x \in \mathbb{R}^C : x \neq 0} \frac{\|Mx\|}{\|x\|} = \frac{1}{\|M^{-1}\|}.$$

Proof of Lemma 22. We begin by rewriting the infimum as the reciprocal of a supremum:

$$\inf_{x \in \mathbb{R}^C : x \neq 0} \frac{\|Mx\|}{\|x\|} = \left(\sup_{x \in \mathbb{R}^C : x \neq 0} \frac{\|x\|}{\|Mx\|} \right)^{-1}.$$

Next, applying the change of variables $x = M^{-1}y$, we have

$$\sup_{x \in \mathbb{R}^C : x \neq 0} \frac{\|x\|}{\|Mx\|} = \sup_{y \in \mathbb{R}^C : y \neq 0} \frac{\|M^{-1}y\|}{\|y\|} = \|M^{-1}\|$$

where the last equality holds by definition. □

Proof of Proposition 6. We are now ready to conclude the proof. Putting it all together, we have

$$
\begin{aligned}
&\mathcal{C}_{\ell_T,P_T,x}(q) - \mathcal{C}^*_{\ell_T,P_T,x} \\
&= \mathtt{KL}(Tp \| Tq) \qquad \because \text{Equation (9)} \\
&\geq \frac{1}{2}\|Tp - Tq\|_1^2 \qquad \because \text{Theorem 20, Pinsker inequality} \\
&= \frac{1}{2}\|T(p-q)\|_1^2 \\
&\geq \frac{1}{2}\frac{\|p-q\|_1^2}{\|T^{-1}\|_1^2} \qquad \because \text{Lemma 22} \\
&\geq \frac{1}{2}\frac{(\underline{L}(q,p) - \underline{L}(p,p))^2}{\|T^{-1}\|_1^2}. \qquad \because \text{Lemma 21} \\
&= \frac{1}{2}\frac{(\mathcal{C}_{L,P,x}(q) - \mathcal{C}^*_{L,P,x})^2}{\|T^{-1}\|_1^2} \qquad \because \text{Equation (12)}
\end{aligned}
$$

as desired. This concludes the proof of Proposition 18. ☐

C Remakrs for Section 4

C.1 Remarks on the Setting of LMNTM

Instead of letting $\left(X, \tilde{Y}\right) \overset{i.i.d.}{\sim} P_{T_i}$, which is a more common assumption, we choose the setting described in Section 4.1 because it fits LLP more naturally. When reducing LLP to LLN, a bag in group i is modeled as a collection of data points sampled from $P_{T_i}(\cdot \mid c)$. If we assume all data points in group i are sampled $i.i.d.$ from P_{T_i}, then we need $(n_{i,1}, n_{i,2}, \ldots, n_{i,C})$, the size of bags in group i, to follow a multinomial distribution, which is too restrictive. Our current setting is more flexible and allows $n_{i,c}$ to be either deterministic or random.

D Proofs for Section 4

D.1 Proof of Theorem 7

Proof of Theorem 7. By Theorem 17, $\forall i \in \mathbb{N}, \exists$ a strictly increasing continuous function θ_i with $\theta_i(0) = 0$ and

$$\theta_i\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,Pg}\big) \leq \mathcal{R}_{\ell_{T_i},P_{T_i}}(f) - \mathcal{R}^*_{\ell_{T_i},P_{T_i}}.$$

Then,

$$\sum_{i=1}^N w_i\theta_i\big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\big) \leq \sum_{i=1}^N w_i\Big(\mathcal{R}_{\ell_{T_i},P_{T_i}}(f) - \mathcal{R}^*_{\ell_{T_i},P_{T_i}}\Big) = \widetilde{\mathcal{R}}_{\ell,P,\mathcal{T}}(f) - \widetilde{\mathcal{R}}^*_{\ell,P,\mathcal{T}}$$

The last equality is implied by the fact $\mathcal{R}^*_{\ell_{T_i},P_{T_i}} = \mathcal{R}_{\ell_{T_i},P_{T_i}}(\eta(x)) < \infty$. Let $\theta = \sum_{i=1}^N w_i\theta_i$ which is clearly continuous and satisfies $\theta(0) = 0$. ☐

D.2 Proof of Theorem 11

Now we introduce a sequence of lemmas to prove the generalization error bound.

Lemma 23. *Let $\mathcal{G} \subset \psi \circ \mathcal{F}$ s.t. $\sup_{x\in X, g\in\mathcal{G}}\|g(x)\|_2 \leq A$ for some constant A. Let $N \in \mathbb{N}$ and $\mathcal{T} = \{T_i\}_{i=1}^N$ be a sequence of invertible column-stochastic matrices. Fix $(w_1, w_2, \ldots, w_N)^{tr} \in \Delta^N$ and $n_{i,c} \in \mathbb{N}$ for each $i \in \mathbb{N}_N$ and $c \in \mathcal{Y}$. Let $S = \left\{X_{i,c,j} : i = \mathbb{N}_N, c \in \mathcal{Y}, j = \mathbb{N}_{n_{i,c}}\right\}$ where each $X_{i,c,j}$ is drawn from the class conditional distribution $P_{T_i}(\cdot \mid c)$ and all $X_{i,c,j}$'s are independent. $\forall i \in \mathbb{N}_N$ and $c \in \mathcal{Y}$, let $\alpha_i \in \mathring{\Delta}^C$ s.t. $\alpha_i(c) = P_{T_i}(\tilde{Y} = c)$. Let ℓ be a proper loss s.t. $\forall i, c$ the function $\lambda_{\ell_{T_i}}(\cdot, c)$ is Lipschitz w.r.t. the 2-norm. Write*

$$\hat{\mathcal{R}}_{w,S}(g) := \sum_{i=1}^N w_i \sum_{c=1}^C \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_i} \lambda_{\ell_{T_i}}(g(X_{i,c,j}), c))$$

and

$$\widetilde{\mathcal{R}}(g) := \widetilde{R}_{\ell,P,\mathcal{T}}\big(\psi^{-1} \circ g\big) = \mathbb{E}\Big[\hat{\mathcal{R}}_{w,S}(g)\Big].$$

Then $\forall \delta \in (0,1]$*, with probability at least* $1-\delta$ *w.r.t. to the draw of S,*

$$\sup_{g\in\mathcal{G}}\Big|\hat{\mathcal{R}}_{w,S}(g) - \widetilde{\mathcal{R}}(g)\Big| \leq \sqrt{2\log\frac{2}{\delta}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2\alpha_i^2(c)}{n_{i,c}}\Big(\big|\lambda_{\ell_{T_i}}\big|A + \big|\lambda_{\ell_{T_i}}\big|_0\Big)^2}$$

$$+ 2\mathbb{E}_{S,\epsilon_{i,c,j}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{N} w_i \sum_{c=1}^{C}\frac{\alpha_i(c)}{n_{i,c}}\sum_{j=1}^{n_{i,c}}\epsilon_{i,c,j}\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c)\right],$$

where $\epsilon_{i,c,j}, i = \mathbb{N}_N, c \in \mathcal{Y}, j \in \mathbb{N}_{n_{i,c}}$ *are i.i.d. Rademacher random variables,* $\big|\lambda_{\ell_{T_i}}\big|_0 = \max_c\big|\lambda_{\ell_{T_i}}(0,c)\big|$*, and* $\big|\lambda_{\ell_{T_i}}\big|$ *is the smallest real number such that it is a Lipschitz constant of* $\lambda_{\ell_{T_i}}(\cdot,c)$*,* $\forall i, c$*.*

Proof. Write

$$\xi(S) := \sup_{g\in\mathcal{G}}\Big|\hat{R}_{w,S}(g) - \widetilde{R}(g)\Big|,$$

$$\xi^+(S) := \sup_{g\in\mathcal{G}}\hat{R}_{w,S}(g) - \widetilde{R}(g), \quad \text{and} \quad \xi^-(S) := \sup_{g\in\mathcal{G}} -\Big(\hat{R}_{w,S}(g) - \widetilde{R}(g)\Big).$$

We will show that the same bound on $\xi^+(S)$ and $\xi^-(S)$ holds with probability at least $1-\frac{\delta}{2}$. Combining these bounds gives the desired bound on $\xi(S)$. We first consider $\xi^+(S)$. The analysis for $\xi^-(S)$ is identical. By definition,

$$\xi^+(S) = \sup_{g\in\mathcal{G}}\sum_{i=1}^{N} w_i\left[\sum_{c=1}^{C}\frac{\alpha_i(c)}{n_{i,c}}\sum_{j=1}^{n_i}\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c) - \mathcal{R}_{\ell_{T_i},P_{T_i}}\big(\psi^{-1}\circ g\big)\right].$$

We first use the Bounded Difference Inequality [26] to bound $\xi^+(S) - \mathbb{E}\xi^+(S)$. Substitute $X_{i,c,j}$ with arbitrary $X'_{i,c,j}$ and $\xi^+(S)$ changes by at most $\sup_{g\in\mathcal{G}}\frac{w_i\alpha_i(c)}{n_{i,c}}\big|\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c) - \lambda_{\ell_{T_i}}\big(g(X'_{i,c,j}),c\big)\big|$. Furthermore,

$$\begin{aligned}\big|\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c)\big| &\leq \big|\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c) - \lambda_{\ell_{T_i}}(0,c)\big| + \big|\lambda_{\ell_{T_i}}(0,c)\big| \\ &\leq \big|\lambda_{\ell_{T_i}}\big|\|g(X_{i,c,j})\| + \big|\lambda_{\ell_{T_i}}\big|_0 \\ &\leq \big|\lambda_{\ell_{T_i}}\big|A + \big|\lambda_{\ell_{T_i}}\big|_0.\end{aligned}$$

Hence,

$$\sup_{g\in\mathcal{G}}\frac{w_i\alpha_i(c)}{n_{i,c}}\big|\lambda_{\ell_{T_i}}(g(X_{i,c,j}),c) - \lambda_{\ell_{T_i}}\big(g(X'_{i,c,j}),c\big)\big| \leq 2\frac{w_i\alpha_i(c)}{n_{i,c}}\Big(\big|\lambda_{\ell_{T_i}}\big|A + \big|\lambda_{\ell_{T_i}}\big|_0\Big).$$

By the Bounded Difference Inequality, with probability at least $1-\frac{\delta}{2}$,

$$\xi^+(S) - \mathbb{E}\xi^+(S) \leq \sqrt{2\log\frac{2}{\delta}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{w_i^2\alpha_i(c)^2}{n_{i,c}}\Big(\big|\lambda_{\ell_{T_i}}\big|A + \big|\lambda_{\ell_{T_i}}\big|_0\Big)^2}.$$

It remains to bound $\mathbb{E}\xi^{+}(S)$. Let $S' = \left\{X'_{i,c,j}, : i = \mathbb{N}_N, c \in \mathcal{Y}, j = \mathbb{N}_{n_i}\right\}$ where every pair of $X'_{i,c,j}$ and $X_{i,c,j}$ are i.i.d. and all $X'_{i,c,j}$'s are independent. Hence,

$$
\begin{align}
\mathbb{E}_S[\xi(S)] &= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \lambda_{\ell_{T_i}}(g(X_{i,c,j}), c) - \sum_{i=1}^{N} w_i \mathcal{R}_{\ell_{T_i}, P_{T_i}}(\psi^{-1} \circ g)\right] \notag \\
&= \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \Big(\hat{\mathcal{R}}_{w,S}(g) - \mathbb{E}_{S'}\hat{\mathcal{R}}_{w,S'}(g)\Big)\right] \notag \\
&\le \mathbb{E}_S\mathbb{E}_{S'}\left[\sup_{g\in\mathcal{G}} \Big(\hat{\mathcal{R}}_{w,S}(g) - \hat{\mathcal{R}}_{w,S'}(g)\Big)\right] \tag{13} \\
&= \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \Big(\lambda_{\ell_{T_i}}(g(X_{i,c,j}), c) - \lambda_{\ell_{T_i}}(g(X'_{i,c,j}), c)\Big)\right] \notag \\
&= \mathbb{E}_{S,S',\epsilon_{i,c,j}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \epsilon_{i,c,j}\Big(\lambda_{\ell_{T_i}}(g(X_{i,c,j}), c) - \lambda_{\ell_{T_i}}(g(X'_{i,c,j}), c)\Big)\right] \tag{14} \\
&\le \mathbb{E}_{S,S',\epsilon_{i,c,j}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \epsilon_{i,c,j}\lambda_{\ell_{T_i}}(g(X_{i,c,j}), c)\right] \notag \\
&\quad + \mathbb{E}_{S,S',\epsilon_{i,c,j}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \epsilon_{i,c,j}\lambda_{\ell_{T_i}}(g(X'_{i,c,j}), c)\right] \tag{15} \\
&= 2\mathbb{E}_{S,\epsilon_{i,c,j}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \epsilon_{i,c,j}\lambda_{\ell_{T_i}}(g(X_{i,c,j}), c)\right]. \notag
\end{align}
$$

(13) is implied by the convexity of $\sup_{g\in\mathcal{G}}$ and Jensen's inequality. The equality in (14) holds because $X'_{i,c,j}$ and $X_{i,c,j}$ are *i.i.d.* and $\epsilon_{i,c,j}$ is symmetric. (15) can be justified by the elementary property of supremum and symmetry of $\epsilon_{i,c,j}$. □

We need the next two lemmas to get rid of the $\lambda_{\ell_{T_i}}$'s when the set $\mathcal{V} \subset \mathbb{R}^C$.

Lemma 24. *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}^C$, let ϕ be a function from $\mathcal{H}$ to $\mathbb{R}$, let a be a positive real number, and let $\lambda : \mathbb{R}^C \to \mathbb{R}$ be a Lipschitz function w.r.t. the norm $\|\cdot\|_2$. We denote the Lipschitz constant of λ by $|\lambda|$. Then,*

$$
\mathbb{E}_{\epsilon} \sup_{f\in\mathcal{H}} \epsilon a \lambda(f(x)) + \phi(f) \le \mathbb{E}_{\epsilon_1,\ldots,\epsilon_C} \sup_{f\in\mathcal{H}} \sqrt{2} a |\lambda| \sum_{c=1}^{C} \epsilon_c f_c(x) + \phi(f)
$$

where $\epsilon, \epsilon_1, \ldots, \epsilon_C$ are independent Rademacher variables and $f_c(x)$ denotes the c-th entry of $f(x)$.

Proof. By Proposition 1 of Maurer [25],

$$
\forall M \in \mathbb{N}, \forall v \in \mathbb{R}^M, \|v\|_2 \le \sqrt{2}\mathbb{E}_{\epsilon_m}\left|\sum_{m=1}^{M} v_m \epsilon_m\right|. \tag{16}
$$

Fix $\delta > 0$, then $\exists f^*, g^* \in \mathcal{F}$,

$$
\begin{aligned}
& 2\left[\mathbb{E}_{\epsilon} \sup_{f\in\mathcal{H}} \epsilon a\lambda(f(x)) + \phi(f)\right] - \delta \\
&= \sup_{f,g\in\mathcal{H}} \left[a\lambda(f(x)) + \phi(f) - a\lambda(g(x)) + \phi(g)\right] - \delta \\
&< a(\lambda(f^*(x)) - \lambda(g^*(x))) + \phi(f^*) + \phi(g^*) \tag{17} \\
&\leq a|\lambda| \|f^*(x) - g^*(x)\|_2 + \phi(f^*) + \phi(g^*) \\
&\leq \mathbb{E}_{\epsilon_c} \sqrt{2} a|\lambda| \left|\sum_{c=1}^{C} \epsilon_c (f^*_c(x) - g^*_c(x))\right| + \phi(f^*) + \phi(g^*) \tag{18} \\
&\leq \mathbb{E}_{\epsilon_c} \sup_{f,g\in\mathcal{H}} \left[\sqrt{2} a|\lambda| \left|\sum_{c=1}^{C} \epsilon_c (f_c(x) - g_c(x))\right| + \phi(f) + \phi(g)\right] \\
&= \mathbb{E}_{\epsilon_c} \sup_{f\in\mathcal{H}} \left[\sqrt{2} a|\lambda| \sum_{c=1}^{C} \epsilon_c f_c(x) + \phi(f)\right] + \mathbb{E}_{\epsilon_c} \sup_{g\in\mathcal{H}} \left[-\sqrt{2} a|\lambda| \sum_{c=1}^{C} \epsilon_c g_c(x) + \phi(g)\right] \tag{19} \\
&= 2\mathbb{E}_{\epsilon_c} \sup_{f\in\mathcal{H}} \left[\sqrt{2} a|\lambda| \sum_{c=1}^{C} \epsilon_c f_c(x) + \phi(f)\right]
\end{aligned}
$$

The existence of f^*, g^* satisfying the inequality in step (17) is guaranteed by the definition of supremum. Step (18) is implied by (16). In (19), we drop the absolute value as we can make $\sum_{c=1}^{C} \epsilon_c (f_c(x) - g_c(x))$ non-negative by exchanging f and g for any realization of $\epsilon_c, c = 1, \ldots, C$. $\square$

Now we move on to the next step.

Lemma 25. *Let $N, C \in \mathbb{N}$. Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}^C$. $\forall i = 1, \ldots, N$, let w_i be a positive real numbers, and let $\lambda_i : \mathbb{R}^C \to \mathbb{R}$ a Lipschitz function. Denote the Lipschitz constant of λ_i by $|\lambda_i|$. Then,*

$$
\mathbb{E}_{\epsilon_i} \sup_{f\in\mathcal{H}} \sum_{i=1}^{N} \epsilon_i w_i \lambda_i(f(x_i)) \leq \sqrt{2} \mathbb{E}_{\epsilon_{i,c}} \sup_{f\in\mathcal{H}} \sum_{i=1}^{N} w_i |\lambda_i| \sum_{c=1}^{C} \epsilon_{i,c} f_c(x_i)
$$

where ϵ_i's and $\epsilon_{i,c}$'s are independent Rademacher variables and $f_c(x)$ denotes the c-th entry of $f(x)$.

Proof. Let $m = 0, 1, \ldots, N$. We prove

$$
\mathbb{E}_{\epsilon_i} \sup_{f\in\mathcal{H}} \sum_{i=1}^{N} \epsilon_i w_i \lambda_i(f(x_i)) \leq
\mathbb{E}_{\epsilon_{i,c},\epsilon_i} \left[\sup_{f\in\mathcal{H}} \sqrt{2} \sum_{1\leq i\leq m} w_i |\lambda_i| \sum_{c=1}^{C} \epsilon_{i,c} f(x_i) + \sum_{m<i\leq N} \epsilon_i w_i \lambda_i(f(x_i))\right]
$$

by induction on m.

The base case when $m = 0$ holds with equality. The case when $m = N$ is the desired inequality. Now, suppose the inequality hold for $m - 1$.

$$
\begin{aligned}
&\mathbb{E}_{\epsilon_i} \sup_{f\in\mathcal{H}} \sum_{i=1}^{N} \epsilon_i w_i \lambda_i(f(x_i)) \\
&\le \mathbb{E}_{\epsilon_{i,c},\epsilon_i}\left[\sup_{f\in\mathcal{H}} \sqrt{2} \sum_{1\le i<m} w_i |\lambda_i| \sum_{c=1}^{C} \epsilon_{i,c} f_c(x_i) + \sum_{m\le i\le N} \epsilon_i w_i \lambda_i(f(x_i))\right] \\
&= \mathbb{E}_{\epsilon_{i,c},\epsilon_i}\left[\sup_{f\in\mathcal{H}} \epsilon_m w_m \lambda_m(f(x_m)) + \phi(f)\right] \\
&= \mathbb{E}_{\{\epsilon_{i,c},\epsilon_i | i\neq m\}} \mathbb{E}_{\epsilon_{m,c}}\left[\sup_{f\in\mathcal{H}} \epsilon_m w_m \lambda_m(f(x_m)) + \phi(f)\right] \\
&\le \mathbb{E}_{\{\epsilon_{i,c},\epsilon_i | i\neq m\}} \mathbb{E}_{\epsilon_{m,c}}\left[\sup_{f\in\mathcal{H}} \sqrt{2} w_m |\lambda_m| \sum_{c=1}^{C} \epsilon_{m,c} f_c(x_m) + \phi(f)\right] \\
&= \mathbb{E}_{\epsilon_{i,c},\epsilon_i}\left[\sup_{f\in\mathcal{H}} \sqrt{2} \sum_{1\le i\le m} w_i |\lambda_i| \sum_{c=1}^{C} \epsilon_{i,c} f_c(x_i) + \sum_{m< i\le N} \epsilon_i w_i \lambda_i(f(x_i))\right]
\end{aligned}
$$

In the first equality, we let $\phi(f)$ denote the rest of the summation. □

Lemma 26. *Let $\mathcal{G} \subset \psi \circ \mathcal{F}$ s.t. $\sup_{x\in X, g\in\mathcal{G}} \|g(x)\|_2 \le A$ for some constant A. Let $N \in \mathbb{N}$ and $\mathcal{T} = \{T_i\}_{i=1}^{N}$ be a sequence of invertible column-stochastic matrices. Fix $(w_1, w_2, \dots, w_N)^{tr} \in \Delta^N$ and $n_{i,c} \in \mathbb{N}$ for each $i \in \mathbb{N}_N$ and $c \in \mathcal{Y}$. Let $S = \left\{X_{i,c,j} : i = \mathbb{N}_N, c \in \mathcal{Y}, j = \mathbb{N}_{n_{i,c}}\right\}$ where each $X_{i,c,j}$ is drawn from the class conditional distribution $P_{T_i}(\cdot \mid c)$ and all $X_{i,c,j}$'s are independent. $\forall i \in \mathbb{N}_N$ and $c \in \mathcal{Y}$, let $\alpha_i \in \mathring{\Delta}^C$ s.t. $\alpha_i(c) = P_{T_i}(\tilde{Y} = c)$. Let ℓ be a proper loss s.t. $\forall i, c$ the function $\lambda_{\ell_{T_i}}(\cdot, c)$ is Lipschitz. Write*

$$
\hat{\mathcal{R}}_{w,S}(g) := \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_i} \lambda_{\ell_{T_i}}(g(X_{i,c,j}), c))
$$

and

$$
\widetilde{\mathcal{R}}(g) := \widetilde{R}_{\ell,P,\mathcal{T}}\left(\psi^{-1} \circ g\right) = \mathbb{E}\left[\hat{\mathcal{R}}_{w,S}(g)\right].
$$

Then $\forall \delta \in (0, 1]$, with probability at least $1 - \delta$ w.r.t. to the draw of S,

$$
\begin{aligned}
\sup_{g\in\mathcal{G}} \left|\hat{\mathcal{R}}_{w,S}(g) - \widetilde{\mathcal{R}}(g)\right| \le &\sqrt{2 \log\frac{2}{\delta} \sum_{i=1}^{N} \sum_{c=1}^{C} \frac{w_i^2 \alpha_i^2(c)}{n_{i,c}} \left(\left|\lambda_{\ell_{T_i}}\right| A + \left|\lambda_{\ell_{T_i}}\right|_0\right)^2} \\
&+ 2\mathbb{E}_{S,\epsilon_{i,c,j,c'}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \left|\lambda_{\ell_{T_i}}\right| \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \sum_{c'=1}^{C} \epsilon_{i,c,j,c'} g_{c'}(X_{i,c,j})\right],
\end{aligned}
$$

where $\epsilon_{i,c,j,c'}, i \in \mathbb{N}_N, c \in \mathcal{Y}, c' \in \mathcal{Y}, j \in \mathbb{N}_{n_{i,c}}$ are i.i.d. Rademacher random variables, $\left|\lambda_{\ell_{T_i}}\right|_0 = \max_c \left|\lambda_{\ell_{T_i}}(0, c)\right|$, and $\left|\lambda_{\ell_{T_i}}\right|$ is the smallest real number such that it is a Lipschitz constant of $\lambda_{\ell_{T_i}}(\cdot, c)$, $\forall i, c$.

Proof of Theorem 26. The theorem is a direct result of Lemmas 23 and 25. □

In Theorem 7, we saw that $\widetilde{\mathcal{R}}(g)$ is a risk for LMNTM satisfying an excess risk bound. Lemma 26 shows that $\hat{\mathcal{R}}_{w,S}(g)$ is an accurate estimate of $\widetilde{\mathcal{R}}(g)$, and therefore justifies its use as an empirical objective for LMNTM.

The second term on the right hand side of the inequality in Lemma 26 depends on the choice of hypothesis class $\mathcal{G}$, and can be viewed as a generalization of Rademacher complexity to LMNTM. To make this term more concrete, we study two popular choices of function classes, the reproducing kernel Hilbert space (RKHS) and the multilayer perceptron (MLP). We first consider the kernel class.

Proposition 27. *Let k be a symmetric positive definite (SPD) kernel, and let $\mathcal{H}$ be the associated reproducing kernel Hilbert space (RKHS). Assume k bounded by K, meaning $\forall x$, $\|k(\cdot,x)\|_{\mathcal{H}} \le K$. Let $\mathcal{G}^k_{K,R}$ denote the ball of radius R in $\mathcal{H}$ and $\mathcal{G} = \underbrace{\mathcal{G}^k_{K,R}\times\mathcal{G}^k_{K,R}\times\cdots\times\mathcal{G}^k_{K,R}}_{\text{the Cartesian product of } C\ \mathcal{G}^k_{K,R}\text{'s}}$. Then*

$$\mathbb{E}_{S,\epsilon_{i,c,j,c'}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{N} w_i\left|\lambda_{\ell_{T_i}}\right|\sum_{c=1}^{C}\frac{\alpha_i(c)}{n_{i,c}}\sum_{j=1}^{n_{i,c}}\sum_{c'=1}^{C}\epsilon_{i,c,j,c'}g_{c'}(X_{i,c,j})\right] \le CRK\sqrt{\sum_{i=1}^{N} w_i^2\left|\lambda_{\ell_{T_i}}\right|^2\sum_{c=1}^{C}\frac{\alpha_i^2(c)}{n_{i,c}}},$$

where $\epsilon_{i,c,j,c'}, i\in\mathbb{N}_N, c\in\mathcal{Y}, c'\in\mathcal{Y}, j\in\mathbb{N}_{n_{i,c}}$ are i.i.d. Rademacher random variables. Thus the generalization error bound becomes: $\forall\delta\in[0,1]$, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left|\hat{\mathcal{R}}_{w,S}(g)-\widetilde{\mathcal{R}}(g)\right| \le \left(\max_i\left(\left|\lambda_{\ell_{T_i}}\right|A+\left|\lambda_{\ell_{T_i}}\right|_0\right)\sqrt{2\log\frac{2}{\delta}}+CRK\max_i\left|\lambda_{\ell_{T_i}}\right|\right)\sqrt{\sum_{i=1}^{N} w_i^2\sum_{c=1}^{C}\frac{\alpha_i^2(c)}{n_{i,c}}}.$$

Proof of Proposition 27. For the reader's convenience, we restate the result:

Proposition 28. *Let k be a symmetric positive definite (SPD) kernel bounded by K and $\mathcal{H}$ be the associated reproducing kernel Hilbert space (RKHS). i.e.$\|k(\cdot,x)\|_{\mathcal{H}}\le K$. Let $\mathcal{G}^k_{K,R}$ denote the ball of radius R in $\mathcal{H}$ and $\mathcal{G} = \underbrace{\mathcal{G}^k_{K,R}\times\mathcal{G}^k_{K,R}\times\cdots\times\mathcal{G}^k_{K,R}}_{\text{the Cartesian products of } C\ \mathcal{G}^k_{K,R}\text{'s}}$. Then*

$$\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{G}^k_{K,R}}\sum_{i=1}^{M}a_i\sum_{c=1}^{C}\epsilon_{i,c}g_c(x_i)\right]\le CRK\sqrt{\sum_{i=1}^{M}a_i^2}$$

where $a_i>0$, and $\epsilon_{i,c}$ are independent Rademacher random variables.

Proof. First, by Cauchy-Schwartz inequality, observe $\forall R>0, g\in\mathcal{G}^k_{K,R}, x\in\mathcal{X}$

$$|g(x)| = |\langle g,k(\cdot,x)\rangle| \le \|g\|_{\mathcal{H}}\|k(\cdot,x)\|_{\mathcal{H}} \le RK.$$

Thus,

$$\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{G}^k_{K,R}}\sum_{i=1}^{M}a_i\sum_{c=1}^{C}\epsilon_{i,c}g_c(x_i)\right]$$

$$=\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{G}^k_{K,R}}\sum_{i=1}^{M}a_i\sum_{c=1}^{C}\epsilon_{i,c}\langle g_c,k(\cdot,x_i)\rangle\right]\tag{20}$$

$$=\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{G}^k_{K,R}}\sum_{c=1}^{C}\langle g_c,\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)\rangle\right]$$

$$=\mathbb{E}_{\epsilon_{i,c}}\left[\sum_{c=1}^{C}\langle R\frac{\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)}{\|\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)\|},\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)\rangle\right]\tag{21}$$

$$=R\sum_{c=1}^{C}\mathbb{E}_{\epsilon_{i,c}}\sqrt{\left\|\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)\right\|^2}$$

$$\leq R\sum_{c=1}^{C}\sqrt{\mathbb{E}_{\epsilon_{i,c}}\left\|\sum_{i=1}^{M}a_i\epsilon_{i,c}k(\cdot,x_i)\right\|^2} \tag{22}$$

$$= CR\sqrt{\sum_{i=1}^{M}a_i^2\left\|k(\cdot,x_i)\right\|^2} \tag{23}$$

$$= CRK\sqrt{\sum_{i=1}^{M}a_i^2}$$

Equality (20) and (21) follow the reproducing property and the equality condition of Cauchy-Schwarz, respectively. (22) is implied by Jensen's inequality and (23) by the independence of Rademacher random variables. □

We now define the Rademacher Complexity-like term $\mathbb{E}_{\epsilon_i}\sup_{g\in\mathcal{G}}\sum_{i=1}^{M}a_i\epsilon_i g(x_i)$ formally and characterize several properties which will be used in the proof of Proposition 31.

Definition 29. *Let $\mathcal{G}$ be a subset of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Denote the sample path $S=(x_i)_{i=1}^{M}$ and weights by $a=(a_i)_{i=1}^{M}$ where $a_i\geq 0$. Define*

$$\mathit{Rad}_{S,a}(\mathcal{G})=\mathbb{E}_{\epsilon_i}\sup_{g\in\mathcal{G}}\sum_{i=1}^{M}a_i\epsilon_i g(x_i),$$

where ϵ_i's are i.i.d. Rademacher random variables.

Proposition 30. *$\mathit{Rad}_{S,a}$ has the following properties:*

1. $\mathcal{G}\subset\mathcal{H}\implies \mathit{Rad}_{S,a}(\mathcal{G})\leq\mathit{Rad}_{S,a}(\mathcal{H})$
2. $\mathit{Rad}_{S,a}(\mathcal{G}_1+\mathcal{G}_2)=\mathit{Rad}_{S,a}(\mathcal{G}_1)+\mathit{Rad}_{S,a}(\mathcal{G}_2)$,
   where $\mathcal{G}_1+\mathcal{G}_2=\{g_1+g_2:g_1\in\mathcal{G}_1,g_2\in\mathcal{G}_2\}$
3. $\forall c_0\in\mathbb{R},\mathit{Rad}_{S,a}(c_0\mathcal{G})=|c_0|\mathit{Rad}_{S,a}(\mathcal{G})$, *where* $c_0\mathcal{G}:=\{c_0g:g\in\mathcal{G}\}$
4. $\mathit{Rad}_{S,a}(\operatorname{conv}\mathcal{G})=\mathit{Rad}_{S,a}(\mathcal{G})$, *where* $\operatorname{conv}\mathcal{G}$ *denotes the convex hull of* $\mathcal{G}$.
5. *Let $\mu:\mathbb{R}\to\mathbb{R}$ be a Lipschitz function and let $|\mu|$ be its Lipschitz constant. Then,*

$$\mathit{Rad}_{S,a}(\mu\circ\mathcal{G})\leq|\mu|\mathit{Rad}_{S,a}(\mathcal{G}),\ \text{where } \mu\circ\mathcal{G}=\{\mu\circ g:g\in\mathcal{G}\}.$$

Proof. Property 1 and 2 immediately follow the definition. Property 3 is implied by the invariance of ϵ_i under negation. It remains to prove Property 4 and 5.

For Property 4:

$$\begin{aligned}
&\mathrm{Rad}_{S,a}(\operatorname{conv}\mathcal{G})\\
&=\mathbb{E}\sup_{n\in\mathbb{N}}\sup_{\lambda\in\Delta^n,g_j\in\mathcal{G}}\sum_{i=1}^{M}a_i\epsilon_i\sum_{j=1}^{n}\lambda_j g_j(x_i)\\
&=\mathbb{E}\sup_{n\in\mathbb{N}}\sup_{\lambda\in\Delta^n,g_j\in\mathcal{G}}\sum_{j=1}^{n}\lambda_j\sum_{i=1}^{M}a_i\epsilon_i g_j(x_i)\\
&=\mathbb{E}\sup_{n\in\mathbb{N}}\sup_{\lambda\in\Delta^n,g_j\in\mathcal{G}}\max_{j}\sum_{i=1}^{M}a_i\epsilon_i g_j(x_i)\\
&=\mathbb{E}\sup_{g\in\mathcal{G}}\sum_{i=1}^{M}a_i\epsilon_i g(x_i)\\
&=\mathrm{Rad}_{S,a}(\mathcal{G}).
\end{aligned}$$

For Property 5, we follow the idea of Meir and Zhang [27],

$$
\begin{aligned}
&\mathrm{Rad}_{S,a}(\mu \circ \mathcal{G}) \\
&= \mathbb{E}_{\epsilon_i} \sup_{g \in \mathcal{G}} \sum_{i=1}^{M} a_i \epsilon_i (\mu \circ g)(x_i) \\
&= \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \mathbb{E}_{\epsilon_1} \sup_{g \in \mathcal{G}} \sum_{i=1}^{M} a_i \epsilon_i (\mu \circ g)(x_i) \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \left[\sup_{g \in \mathcal{G}} \left(a_1 (\mu \circ g)(x_1) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ g)(x_i) \right) \right. \\
&\left. + \sup_{g' \in \mathcal{G}} \left(-a_1 (\mu \circ g')(x_1) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ g')(x_i) \right) \right] \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \left[\sup_{g, g' \in \mathcal{G}} a_1 (\mu(g(x_1)) - \mu(g'(x_1))) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ (g + g'))(x_i) \right] \\
&\le \frac{1}{2} \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \left[\sup_{g, g' \in \mathcal{G}} a_1 |\mu| |g(x_1) - g'(x_1)| + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ (g + g'))(x_i) \right] \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \left[\sup_{g, g' \in \mathcal{G}} a_1 |\mu| (g(x_1) - g'(x_1)) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ (g + g'))(x_i) \right] \qquad (24) \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_i, i=2,3,\ldots,M} \left[\sup_{g \in \mathcal{G}} \left(a_1 |\mu| g(x_1) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ g)(x_i) \right) \right. \\
&\left. + \sup_{g' \in \mathcal{G}} \left(-a_1 |\mu| g'(x_1) + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ g')(x_i) \right) \right] \\
&= \mathbb{E}_{\epsilon_i} \sup_{g \in \mathcal{G}} \left[a_1 |\mu| g(x_1) \epsilon_1 + \sum_{i=2}^{M} a_i \epsilon_i (\mu \circ g)(x_i) \right].
\end{aligned}
$$

In step (24), we can drop the absolute value since we can always make $(g(x_1) - g'(x_1))$ non-negative by exchanging g and g' while leaving the rest of the equation invariant. Proceeding by the above argument inductively on i, we eventually have

$$\mathrm{Rad}_{S,a}(\mu \circ \mathcal{G}) \le \mathbb{E}_{\epsilon_i} \sup_{g \in \mathcal{G}} \sum_{i=1}^{M} a_i |\mu| g(x_i) \epsilon_i = |\mu| \mathrm{Rad}_{S,a}(\mathcal{G})$$

as desired. □

To simplify the notations, we follow Zhang et al. [54] and define the real-valued MLP inductively:

$$\mathcal{N}_1 = \left\{ x \to \langle x, v \rangle : v \in \mathbb{R}^d, \|v\|_2 \le \beta \right\},$$

$$\mathcal{N}_m = \left\{ x \to \sum_{j=1}^{d} w_j \mu(f_j(x)) : v \in \mathbb{R}^d, \|v\|_1 \le \beta, f_j \in \mathcal{N}_{m-1} \right\},$$

where $\beta \in \mathbb{R}_+$ and μ is a 1-Lipschitz activation function. Define an MLP which outputs a vector in $\mathbb{R}^C$ by $\mathcal{G} = \underbrace{\mathcal{N}_m \times \mathcal{N}_m \times \cdots \times \mathcal{N}_m}_{\text{the Cartesian product of } C\ \mathcal{N}_m\text{'s}}$. To leverage standard techniques for the proof, we additionally assume $\forall m \in \mathbb{N}, 0 \in \mu \circ \mathcal{N}_m$.

Proposition 31. *Let* $\mathcal{G} = \underbrace{\mathcal{N}_m \times \mathcal{N}_m \times \cdots \times \mathcal{N}_m}_{\text{the Cartesian product of } C\ \mathcal{N}_m\text{'s}}$. *Assume* $\forall x \in \mathcal{X}, \|x_i\| \leq \alpha$ *and* $\forall m \in \mathbb{N}, 0 \in \mu \circ \mathcal{N}_m$. *Then,*

$$\mathbb{E}_{S,\epsilon_{i,c,j,c'}}\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} w_i \left|\lambda_{\ell_{T_i}}\right| \sum_{c=1}^{C} \frac{\alpha_i(c)}{n_{i,c}} \sum_{j=1}^{n_{i,c}} \sum_{c'=1}^{C} \epsilon_{i,c,j,c'} g_{c'}(X_{i,c,j})\right] \leq C\alpha 2^{m-1}\beta^m \sqrt{\sum_{i=1}^{N} w_i^2 \left|\lambda_{\ell_{T_i}}\right|^2 \sum_{c=1}^{C} \frac{\alpha_i^2(c)}{n_{i,c}}},$$

where $\epsilon_{i,c,j,c'}, i \in \mathbb{N}_N, c \in \mathcal{Y}, c' \in \mathcal{Y}, j \in \mathbb{N}_{n_{i,c}}$ *are i.i.d. Rademacher random variables. Thus, the generalization error bound becomes:* $\forall \delta \in [0,1]$, *with probability at least* $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left|\hat{\mathcal{R}}_{w,S}(g) - \widetilde{\mathcal{R}}(g)\right| \leq \left(\max_i\left(\left|\lambda_{\ell_{T_i}}\right| A + \left|\lambda_{\ell_{T_i}}\right|_0\right)\sqrt{2\log\frac{2}{\delta}} + C\alpha 2^{m-1}\beta^m \max_i\left|\lambda_{\ell_{T_i}}\right|\right)\sqrt{\sum_{i=1}^{N} w_i^2 \sum_{c=1}^{C}\frac{\alpha_i^2(c)}{n_{i,c}}}.$$

Proof of Proposition [31]. For the reader's convenience, we restate the result:

Proposition 32. *Let* $\mathcal{G} = \underbrace{\mathcal{N}_m \times \mathcal{N}_m \times \cdots \times \mathcal{N}_m}_{\text{the Cartesian products of } C\ \mathcal{N}_m\text{'s}}$. *Assume* $\forall x \in \mathcal{X}, \|x_i\| \leq \alpha$ *and* $\forall k \in \mathbb{N}, 0 \in \mu \circ \mathcal{N}_k$. *Then,*

$$\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{N}_m} \sum_{i=1}^{M} a_i \sum_{c=1}^{C} \epsilon_{i,c} g_c(x_i)\right] \leq C\alpha 2^{m-1}\beta^m \sqrt{\sum_{i=1}^{M} a_i^2}.$$

where $a_i > 0$, *and* $\epsilon_{i,c}$ *are independent Rademacher random variables.*

Recall that the MLP outputs a vector in $\mathbb{R}^C$. The set of MLPs is $\mathcal{G} = \underbrace{\mathcal{N}_m \times \mathcal{N}_m \times \cdots \times \mathcal{N}_m}_{\text{the Cartesian products of } C\ \mathcal{N}_m\text{'s}}$ where the set $\mathcal{N}_m$ is defined inductively as

$$\mathcal{N}_1 = \left\{x \to \langle x, v\rangle : v \in \mathbb{R}^d, \|v\|_2 \leq \beta\right\} \quad \text{for } m = 1, \text{ and}$$

$$\mathcal{N}_m = \left\{x \to \sum_{j=1}^{d} w_j \mu(f_j(x)) : v \in \mathbb{R}^d, \|v\|_1 \leq \beta, f_j \in \mathcal{N}_{m-1}\right\} \quad \text{for } m > 1.$$

$\beta \in \mathbb{R}_+$, and μ is a 1-Lipschitz activation function. We now proceed with the proof of Proposition [32].

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{N}_m} \sum_{i=1}^{M} a_i \sum_{c=1}^{C} \epsilon_{i,c} g_c(x_i)\right] \\
&= \mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{N}_m} \sum_{c=1}^{C}\sum_{i=1}^{M} a_i \epsilon_{i,c} g_c(x_i)\right] \\
&\leq \sum_{c=1}^{C} \mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c\in\mathcal{N}_m} \sum_{i=1}^{M} a_i \epsilon_{i,c} g_c(x_i)\right] \\
&= C\mathbb{E}_{\epsilon_i}\left[\sup_{g\in\mathcal{N}_m} \sum_{i=1}^{M} a_i \epsilon_i g(x_i)\right] \\
&= C\mathrm{Rad}_{S,a}(\mathcal{N}_m)
\end{aligned}$$

where

$$\mathrm{Rad}_{S,a}(\mathcal{N}_m) = \mathbb{E}_{\epsilon_i}\left[\sup_{h_j \in \mathcal{N}_{m-1}, \|v\|_1 \le \beta} \sum_{i=1}^{M} a_i \epsilon_i \left(\sum_{j=1}^{d} v_j (\mu \circ h_j)\right)(x_i)\right].$$

Note $\sum_{j=1}^{d} v_j(\mu \circ h_j) \in \beta \operatorname{conv}(\mu \circ \mathcal{N}_{m-1} - \mu \circ \mathcal{N}_{m-1})$. Here the difference between two sets of functions is $\mathcal{G}_1 - \mathcal{G}_2 = \{g_1 - g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$ and $\beta\mathcal{G}_1 = \{\beta g_1 : g_1 \in \mathcal{G}_1\}$ for a real number β. Apply Proposition 30,

$$\begin{aligned}
&\mathrm{Rad}_{S,a}(\mathcal{N}_m) \\
&\le \mathrm{Rad}_{S,a}(\beta \operatorname{conv}(\mu \circ \mathcal{N}_{m-1} - \mu \circ \mathcal{N}_{m-1})) \\
&= \beta \mathrm{Rad}_{S,a}(\operatorname{conv}(\mu \circ \mathcal{N}_{m-1} - \mu \circ \mathcal{N}_{m-1})) \\
&= \beta \mathrm{Rad}_{S,a}((\mu \circ \mathcal{N}_{m-1} - \mu \circ \mathcal{N}_{m-1})) \\
&= \beta(\mathrm{Rad}_{S,a}(\mu \circ \mathcal{N}_{m-1}) + \mathrm{Rad}_{S,a}(-\mu \circ \mathcal{N}_{m-1})) \\
&= 2\beta \mathrm{Rad}_{S,a}(\mu \circ \mathcal{N}_{m-1}) \\
&\le 2|\mu|\beta \mathrm{Rad}_{S,a}(\mathcal{N}_{m-1})
\end{aligned}$$

Proceeding backward inductively on m, we have $\mathrm{Rad}_{S,a}(\mathcal{N}_m) \le 2^{m-1}\beta^{m-1}\mathrm{Rad}_{S,a}(\mathcal{N}_1)$. The set $\mathcal{N}_1$ can be viewed as the ball with radius β centered at 0 in the RKHS associated to linear kernel bounded α, so we can apply Proposition 28. Therefore,

$$\mathrm{Rad}_{S,a}(\mathcal{N}_m) \le 2^{m-1}\beta^{m-1}\mathrm{Rad}_{S,a}(\mathcal{N}_1) \le 2^{m-1}\beta^m \alpha \sqrt{\sum_{i=1}^{M} a_i^2}$$

and

$$\mathbb{E}_{\epsilon_{i,c}}\left[\sup_{g_c \in \mathcal{N}_m} \sum_{i=1}^{M} a_i \sum_{c=1}^{C} \epsilon_{i,c} g_c(x_i)\right] \le C\mathrm{Rad}_{S,a}(\mathcal{N}_m) \le C\alpha 2^{m-1}\beta^m \sqrt{\sum_{i=1}^{M} a_i^2}$$

as desired. ☐

Proof of Theorem 11. Theorem 11 follows Lemma 26, Proposition 27, Proposition 31, and the fact that $\alpha_i(c) \le 1$. ☐

D.3 Proof of Proposition 12

Proof of Proposition 12. By Corollary 1.42 of Weaver [48], $\| \|\nabla_s \lambda_{\ell_T}(s,y)\|_2 \|_\infty$ is a Lipschitz constant of $\lambda_{\ell_T}(\cdot, y)$, where $y \in \{1, 2 \ldots, C\}$, ∇ denotes the gradient of a function, $\|\nabla_s \lambda_{\ell_T}(s,y)\|_2$ is a function maps s to a real number, and the $\|\cdot\|_\infty$ takes the essential supremum over Δ^C. We use $t_{i,j}$ to denote the element at i-row and j-column of T.

$$\lambda_{\ell_T}(s,y) = -\log\left(\sum_{k=1}^{C} t_{y,k} \frac{e^{s_k}}{\sum_{j=1}^{C} e^{s_j}}\right) = -\log\left(\sum_{k=1}^{C} t_{y,k} e^{s_k}\right) + \log\left(\sum_{j=1}^{C} e^{s_j}\right).$$

$$\begin{aligned}
\frac{\partial \lambda_{\ell_T}(s,y)}{\partial s_i} &= -\frac{t_{y,i} e^{s_i}}{\sum_{j=1}^{C} t_{y,j} e^{s_j}} + \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}} = -\frac{t_{y,i} \frac{e^{s_i}}{\sum_{k=1}^{C} e^{s_k}}}{\sum_{j=1}^{C} t_{y,j} \frac{e^{s_j}}{\sum_{j=k}^{C} e^{s_k}}} + \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}} \\
&= -\frac{t_{y,i} p_i}{\sum_{j=1}^{C} t_{y,j} p_j} + p_i
\end{aligned}$$

In the last equality, we denote $\frac{e^{s_i}}{\sum_{k=1}^{C} e^{s_k}}$ by p_i. Then,

$$\begin{aligned} \|\nabla_s \lambda_{\ell_T}(s,y)\|_2^2 &= \sum_{i=1}^{C}\left(-\frac{t_{y,i}p_i}{\sum_{j=1}^{C} t_{y,j}p_j} + p_i\right)^2 \le \sum_{i=1}^{C}\left|-\frac{t_{y,i}p_i}{\sum_{j=1}^{C} t_{y,j}p_j} + p_i\right| \\ &\le \sum_{i=1}^{C}\left(\frac{t_{y,i}p_i}{\sum_{j=1}^{C} t_{y,j}p_j} + p_i\right) = 2 \end{aligned} \tag{25}$$

The inequality in step (25) follows the observation that $\left|-\frac{t_{y,i}p_i}{\sum_{j=1}^{C} t_{y,j}p_j} + p_i\right| \le 1$ □

E Confirmation of Probabilistic Model

In Section 5.2, we state that $\alpha(i) = \bar{P}_T(\tilde{Y} = i)$, $P_{\gamma_i}(\cdot) = \bar{P}_T(\cdot \mid \tilde{Y} = i)$, and $\gamma_i(c) = \bar{P}_T(Y = c \mid \tilde{Y} = i)$ for matrix T with $T(i,j) = \frac{\gamma_i(j)\alpha(i)}{\sigma(j)}$. Here we confirm these facts.

Let T be a stochastic matrix with entries $T(i,j) = \frac{\gamma_i(j)\alpha(i)}{\sigma(j)}$. We construct the joint probability measure $\bar{P}_T$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ as described in Section 2. We can see $\bar{P}_T\left(\tilde{Y} = i\right) = \sum_{j=1}^{C} \bar{P}_T\left(\tilde{Y} = i, Y = j\right) = \sum_{j=1}^{C} \bar{P}_T(Y = j)T(i,j) = \sum_{j=1}^{C} \sigma(j)\frac{\gamma_i(j)\alpha(i)}{\sigma(j)} = \alpha(i)$ and $\forall$ events $\mathcal{A} \subset \mathcal{X}, \forall i, y \in \mathcal{Y}$ $\bar{P}_T\left(\tilde{Y} = i\right) = \alpha_i$ and $\forall \mathcal{A} \in \mathcal{M}_{\mathcal{X}}, \forall i, y \in \mathcal{Y}$

$$\begin{aligned} &\bar{P}_T\left(X \in \mathcal{A}, Y = y \mid \tilde{Y} = i\right) \\ &= \frac{1}{\alpha(i)}\bar{P}_T\left(X \in \mathcal{A}, Y = y, \tilde{Y} = i\right) \\ &= \frac{1}{\alpha(i)}P(X \in \mathcal{A}, Y = y)\frac{\gamma_i(y)\alpha(i)}{\sigma(y)} \\ &= P_y(X \in \mathcal{A})\gamma_i(y) \\ &= P_{\gamma_i}(X \in \mathcal{A}, Y = y). \end{aligned}$$

Hence, $\bar{P}_T\left(\cdot \mid \tilde{Y} = i\right) = P_{\gamma_i}(\cdot)$, which implies that $\bar{P}_T\left(Y = c \mid \tilde{Y} = i\right) = P_{\gamma_i}(Y = c) = \gamma_i(c)$, and for a data point $\left(X, Y, \tilde{Y}\right) \sim \bar{P}_T$ the event $\tilde{Y} = i$ entails that $(X, Y) \sim P_{\gamma_i}$.

F Grouping and Weights Optimization

To optimize the weights or the assignment of bags we would need to optimize the composition of our two bounds: $\theta(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}) \le$ Emprical Risk + Generalization Error Bound $- \mathcal{R}^*_{l,P,\mathcal{T}}$. This is in contrast to the approach with backward correction [39] which does not require the excess risk bound (because their excess target risk is simply proportional to the excess surrogate risk). Therefore, to optimize the composition of our bounds, we'd need to estimate the surrogate Bayes risk, a challenging task. We also note that both the generalization error bound and excess risk bound involve weights w_i and noise matrices T_i. Therefore, even if the surrogate Bayes risk were somehow known, the resulting integer programming problem is much more involved than for the backward correction, where it's a simple matching problem.

Fortunately, LLPFC with random partitioning and weights which optimize solely generalization error bound yields superior empirical results in the experiments and outperforms other multiclass LLP methods by a significant margin. We believe weight optimization is much more important for the backward correction, where the loss functions can have large and disparate magnitudes (which need to be offset by carefully chosen weights), than it is for forward correction where the outputs of the inverse link function are in the unit simplex and thus all of a comparable magnitude. A similar point is made by Patrini et al. [30] in the last two sentence in the first paragraph of section 6.