SUPPLEMENTARY MATERIAL FOR DISTRIBUTIONALLY ROBUST MULTICLASS CLASSIFICATION AND APPLICATIONS IN DEEP IMAGE CLASSIFIERS

Ruidi Chen⋆ *Boran Hao*† *Ioannis Ch. Paschalidis*†

⋆Amazon SCOT
†Department of Electrical and Computer Engineering, Boston University

1. OMITTED RESULTS AND PROOFS

1.1. Omitted Corollaries

The following results are needed to establish Theorem 2.1 in the main paper.

Corollary 1.1. *Define the convex log-loss in class k as $h_{\mathbf{B}}(\mathbf{x}, \mathbf{e}_k) = \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} - (\mathbf{B}\mathbf{e}_k)'\mathbf{x}$. We have:*

$$\sup_{\mathbf{x} \in \mathbb{R}^p} h_{\mathbf{B}}(\mathbf{x}, \mathbf{e}_k) - \lambda \|\mathbf{x} - \mathbf{x}_i\|_r = \begin{cases} h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{e}_k), & \text{if } \lambda \geq \kappa, \\ +\infty, & \text{otherwise}, \end{cases}$$

where $\kappa \triangleq \sup\{\|\mathbf{B}(\boldsymbol{\gamma} - \mathbf{e}_k)\|_s : \boldsymbol{\gamma} \geq \mathbf{0},\ \mathbf{1}'\boldsymbol{\gamma} = 1\}$, with $\mathbf{e}_k$ the k-th unit vector, and $r, s \geq 1$, $1/r + 1/s = 1$.

Proof. The proof of Corollary 1.1 uses the following result, which comes from the proof of Theorem 6.3 in [1].

Corollary 1.2 ([1], Theorem 6.3)**.** *Suppose the loss function $h(\mathbf{x})$ is convex in $\mathbf{x} \in \mathbb{R}^p$. We have:*

$$\sup_{\mathbf{x} \in \mathbb{R}^p} h(\mathbf{x}) - \lambda \|\mathbf{x} - \mathbf{x}_i\|_r = \begin{cases} h(\mathbf{x}_i), & \text{if } \lambda \geq \kappa, \\ +\infty, & \text{otherwise}, \end{cases}$$

where $\kappa \triangleq \sup\{\|\boldsymbol{\theta}\|_s : h^*(\boldsymbol{\theta}) < \infty\}$, $r, s \geq 1$, $1/r + 1/s = 1$, and $h^*(\boldsymbol{\theta})$ denotes the convex conjugate function of $h(\mathbf{x})$.

To prove Corollary 1.1, the key is to compute the value of κ. We define

$$h_k(\mathbf{x}) \triangleq h_{\mathbf{B}}(\mathbf{x}, \mathbf{e}_k) = \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}} - \mathbf{w}_k'\mathbf{x}.$$

The function $h_k(\mathbf{x})$ is convex in $\mathbf{x}$, due to the convexity of $\log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}}$. From Corollary 1.2 we see that, in order to compute κ, we need to find the convex conjugate of $h_k(\mathbf{x})$. To do so, we first compute the convex conjugate of $f(\mathbf{x}) \triangleq \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}}$.

$$\begin{aligned} f^*(\boldsymbol{\theta}) &\triangleq \sup\nolimits_{\mathbf{x} \in \mathbb{R}^p} \{\boldsymbol{\theta}'\mathbf{x} - f(\mathbf{x})\} \\ &= \sup\nolimits_{\mathbf{x} \in \mathbb{R}^p} \{\boldsymbol{\theta}'\mathbf{x} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}}\}. \end{aligned} \tag{1}$$

Write the first-order condition of Problem (1) as:

$$\boldsymbol{\theta} - \frac{\mathbf{B} e^{\mathbf{B}'\mathbf{x}^*}}{\mathbf{1}' e^{\mathbf{B}'\mathbf{x}^*}} = \mathbf{0},$$

where $\mathbf{x}^*$ is the stationary point. This implies that

$$\boldsymbol{\theta} = \mathbf{B}\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} = e^{\mathbf{B}'\mathbf{x}^*} / \mathbf{1}' e^{\mathbf{B}'\mathbf{x}^*}$. We thus have that:

$$f^*(\boldsymbol{\theta}) < \infty, \text{ if } \boldsymbol{\theta} = \mathbf{B}\boldsymbol{\gamma}, \text{ where } \boldsymbol{\gamma} \geq \mathbf{0},\ \mathbf{1}'\boldsymbol{\gamma} = 1.$$

The convex conjugate of $h_k(\mathbf{x})$ can be expressed as:

$$\begin{aligned} h_k^*(\boldsymbol{\theta}) &\triangleq \sup_{\mathbf{x} \in \mathbb{R}^p} \{\boldsymbol{\theta}'\mathbf{x} - h_k(\mathbf{x})\} \\ &= \sup_{\mathbf{x} \in \mathbb{R}^p} \{(\boldsymbol{\theta} + \mathbf{w}_k)'\mathbf{x} - \log \mathbf{1}' e^{\mathbf{B}'\mathbf{x}}\} \\ &= f^*(\boldsymbol{\theta} + \mathbf{w}_k). \end{aligned} \tag{2}$$

To make $h_k^*(\boldsymbol{\theta}) < \infty$, it must satisfy that $\boldsymbol{\theta} + \mathbf{w}_k = \mathbf{B}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \geq \mathbf{0},\ \mathbf{1}'\boldsymbol{\gamma} = 1$. Therefore,

$$\begin{aligned} \kappa &\triangleq \sup\{\|\boldsymbol{\theta}\|_s : h_k^*(\boldsymbol{\theta}) < \infty\} \\ &= \sup\{\|\mathbf{B}\boldsymbol{\gamma} - \mathbf{w}_k\|_s : \boldsymbol{\gamma} \geq \mathbf{0},\ \mathbf{1}'\boldsymbol{\gamma} = 1\}. \end{aligned}$$

□

1.2. Omitted Proof to Theorem 2.1

Proof. Let us first examine the inner supremum of the DRO problem, which can be expressed as:

$$\sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}[h_{\mathbf{B}}(\mathbf{x}, \mathbf{y})] = \sup_{\mathbb{Q} \in \Omega} \int_{\mathcal{Z}} h_{\mathbf{B}}(\mathbf{z}) \mathrm{d}\mathbb{Q}(\mathbf{z}), \tag{3}$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. By definition of the Wasserstein set we can reformulate (3) as:

$$\begin{aligned} \sup_{\Pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} & \int_{\mathcal{Z}} h_{\mathbf{B}}(\mathbf{z}) \mathrm{d}\Pi(\mathbf{z}, \mathcal{Z}) \\ \text{s.t.} & \int_{\mathcal{Z} \times \mathcal{Z}} l(\mathbf{z}, \tilde{\mathbf{z}}) \mathrm{d}\Pi(\mathbf{z}, \tilde{\mathbf{z}}) \leq \epsilon, \\ & \int_{\mathcal{Z} \times \mathcal{Z}} \delta_{\mathbf{z}_i}(\tilde{\mathbf{z}}) \mathrm{d}\Pi(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{1}{N},\ \forall i \in [\![N]\!], \end{aligned} \tag{4}$$

where Π is the joint distribution of $\mathbf{z}$ and $\tilde{\mathbf{z}}$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, $\tilde{\mathbf{z}}$ indexes the support of $\hat{\mathbb{P}}_N$, and $\delta_{\mathbf{z}_i}(\cdot)$ is the Dirac delta function at point $\mathbf{z}_i$. Using $\mathbb{Q}^i$ to denote the conditional distribution of $\mathbf{z}$ given $\tilde{\mathbf{z}} = \mathbf{z}_i$, we can rewrite (4) as:

$$
\begin{aligned}
\sup_{\mathbb{Q}^i} \; & \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{Z}} h_{\mathbf{B}}(\mathbf{z})\mathrm{d}\mathbb{Q}^i(\mathbf{z}) \\
\text{s.t.} \; & \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{Z}} l(\mathbf{z},\tilde{\mathbf{z}})\mathrm{d}\mathbb{Q}^i(\mathbf{z}) \le \epsilon, \\
& \int_{\mathcal{Z}} \mathrm{d}\mathbb{Q}^i(\mathbf{z}) = 1, \; \forall i \in [\![N]\!].
\end{aligned} \tag{5}
$$

Notice that the support $\mathcal{Z}$ can be decomposed into $\mathbb{R}^p$ and a discrete set $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. We thus decompose each distribution $\mathbb{Q}^i$ into unnormalized measures $\mathbb{Q}^i_k$ supported on $\mathbb{R}^p$ such that $\mathbb{Q}^i_k(\mathrm{d}\mathbf{x}) \triangleq \mathbb{Q}^i(\mathrm{d}\mathbf{x}, \mathbf{y} = \mathbf{e}_k), k \in [\![K]\!]$. Problem (5) can then be reformulated as:

$$
\begin{aligned}
\sup_{\mathbb{Q}^i_k} \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\int_{\mathbb{R}^p} h_{\mathbf{B}}(\mathbf{x},\mathbf{e}_k)\mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) \\
\text{s.t.} \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\int_{\mathbb{R}^p} l((\mathbf{x},\mathbf{e}_k),(\mathbf{x}_i,\mathbf{y}_i))\mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) \le \epsilon, \\
& \sum_{k=1}^{K}\int_{\mathbb{R}^p} \mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) = 1, \; \forall i \in [\![N]\!].
\end{aligned} \tag{6}
$$

Using the definition of l, we can write Problem (6) as:

$$
\begin{aligned}
\sup_{\mathbb{Q}^i_k} \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\int_{\mathbb{R}^p} h_{\mathbf{B}}(\mathbf{x},\mathbf{e}_k)\mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) \\
\text{s.t.} \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\int_{\mathbb{R}^p} \Big(\|\mathbf{x}-\mathbf{x}_i\|_r + M\|\mathbf{e}_k - \mathbf{y}_i\|_t\Big)\mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) \le \epsilon, \\
& \sum_{k=1}^{K}\int_{\mathbb{R}^p} \mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) = 1, \; \forall i \in [\![N]\!],
\end{aligned}
$$

which can be equivalently written as:

$$
\begin{aligned}
\sup_{\mathbb{Q}^i_k} \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\int_{\mathbb{R}^p} h_{\mathbf{B}}(\mathbf{x},\mathbf{e}_k)\mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) \\
\text{s.t.} \; & \frac{1}{N}\int_{\mathbb{R}^p}\Big(\sum_{i=1}^{N}\|\mathbf{x}-\mathbf{x}_i\|_r\Big(\sum_{k=1}^{K}\mathrm{d}\mathbb{Q}^i_k(\mathbf{x})\Big) \\
& + 2^{1/t}M\Big(\sum_{k}\sum_{i:\mathbf{y}_i=\mathbf{e}_k}\mathrm{d}\overline{\mathbb{Q}}^i_k(\mathbf{x})\Big)\Big) \le \epsilon, \\
& \sum_{k}\int_{\mathbb{R}^p} \mathrm{d}\mathbb{Q}^i_k(\mathbf{x}) = 1, \; \forall i \in [\![N]\!],
\end{aligned} \tag{7}
$$

where $\overline{\mathbb{Q}}^i_k \triangleq \sum_{j=1}^{K}\mathbb{Q}^i_j - \mathbb{Q}^i_k$. In the derivation we used the fact that $\|\mathbf{e}_i - \mathbf{e}_j\|_t = 2^{1/t}$, if $i \neq j$.

Notice that (7) is a linear problem (LP) in $\mathbb{Q}^i_k$. We can apply linear duality with dual variables λ and s_i. (7) is a special LP in that the decision variables $\mathbb{Q}^i_k$ are infinite dimensional. For each $\mathbb{Q}^i_k(\mathbf{x})$, its coefficients in the constraints of (7) are multiplied by the corresponding dual variables to produce the constraints of the dual problem (8) (LP duality). Since $\mathbb{Q}^i_k$ has infinitely many arguments $\mathbf{x}$, the constraints of (8) involve the supremum over $\mathbf{x}$. The dual problem of (7) can be written as:

$$
\begin{aligned}
\inf_{\lambda\ge 0, s_i} \; & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} \; & \sup_{\mathbf{x}\in\mathbb{R}^p} h_{\mathbf{B}}(\mathbf{x},\mathbf{e}_k) - \lambda\|\mathbf{x}-\mathbf{x}_i\|_r - \lambda M\|\mathbf{e}_k - \mathbf{y}_i\|_t \le s_i, \\
& \forall i \in [\![N]\!], k \in [\![K]\!].
\end{aligned} \tag{8}
$$

Note that the value of Problem 8 is equal to the value of 7 for the optimal dual variable λ, due to strong duality. Using Corollary 1.1, we can write Problem (8) as:

$$
\begin{aligned}
\inf_{\lambda, s_i} \; & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} \; & h_k(\mathbf{x}_i) - \lambda M\|\mathbf{e}_k - \mathbf{y}_i\|_t \le s_i, \; \forall i \in [\![N]\!], k \in [\![K]\!], \\
& \lambda \ge \sup\{\|\mathbf{B}(\boldsymbol{\gamma} - \mathbf{e}_k)\|_s : \boldsymbol{\gamma} \ge \mathbf{0}, \; \mathbf{1}'\boldsymbol{\gamma} = 1\}, \; \forall k \in [\![K]\!],
\end{aligned} \tag{9}
$$

where $1/r + 1/s = 1$. As $M \to \infty$, i.e., we assign a very large weight on the labels, implying that samples from different classes are infinitely far away, the first set of constraints in Problem (9) reduces to: $h_B(\mathbf{x}_i, \mathbf{y}_i) \le s_i, \; \forall i \in [\![N]\!]$. Therefore, the optimal value of (9) is:

$$
\frac{1}{N}\sum_{i=1}^{N} h_B(\mathbf{x}_i, \mathbf{y}_i) + \lambda\epsilon, \tag{10}
$$

where $\lambda = \max_k \sup\{\|\mathbf{B}(\boldsymbol{\gamma} - \mathbf{e}_k)\|_s : \boldsymbol{\gamma} \ge \mathbf{0}, \; \mathbf{1}'\boldsymbol{\gamma} = 1\}$. Note that by setting $M \to \infty$, it does not imply that we only care about the perturbation on the labels; instead, when the samples are in the same class, we focus on perturbations on the input feature $\mathbf{x}$. To compute λ, notice that

$$
\|\mathbf{B}(\boldsymbol{\gamma} - \mathbf{e}_k)\|_s \le \|\mathbf{B}\|_s\|\boldsymbol{\gamma} - \mathbf{e}_k\|_s,
$$

where $\|\mathbf{B}\|_s$ is the induced ℓ_s norm of the matrix $\mathbf{B}$. The maximum of $\|\boldsymbol{\gamma} - \mathbf{e}_k\|_s$ can be obtained as:

$$
\begin{aligned}
\|\boldsymbol{\gamma} - \mathbf{e}_k\|_s^s &= \sum\nolimits_{i=1}^{k-1}\gamma_i^s + (1-\gamma_k)^s + \sum\nolimits_{j=k+1}^{K}\gamma_j^s \\
&\le \sum\nolimits_{i=1}^{k-1}\gamma_i + (1-\gamma_k) + \sum\nolimits_{j=k+1}^{K}\gamma_j \\
&= 1 - 2\gamma_k + 1 \\
&\le 2,
\end{aligned}
$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$. Therefore, (10) can be reformulated into Eq. (5) in the statement of Theorem 2.1 in the main paper by setting $\lambda = 2^{1/s}\|\mathbf{B}\|_s$. $\square$

1.3. Omitted Proof to Theorem 3.2

Proof. We need to find an upper bound to the loss $h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) = \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}$. To do this, we first bound $|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{x}_0,\mathbf{y})|$ for any $\mathbf{x}_0 \in \mathbb{R}^p$. Note that

$$\begin{aligned}
&|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{x}_0,\mathbf{y})| \\
=& |\log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x} - \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}_0} + \mathbf{y}'\mathbf{B}'\mathbf{x}_0| \\
\leq& |\log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}_0}| + |\mathbf{y}'\mathbf{B}'\mathbf{x} - \mathbf{y}'\mathbf{B}'\mathbf{x}_0|.
\end{aligned} \tag{11}$$

Let us examine the two terms in (11) separately. For the first term, define a function $g(\mathbf{a}) = \log \mathbf{1}'e^{\mathbf{a}}$, where $\mathbf{a} \in \mathbb{R}^K$. Using the mean value theorem, we know for any $\mathbf{a},\mathbf{b} \in \mathbb{R}^K$, there exists some $t \in (0,1)$ such that

$$\begin{aligned}
|g(\mathbf{b}) - g(\mathbf{a})| \leq& \big\|\nabla g\big((1-t)\mathbf{a} + t\mathbf{b}\big)\big\|_r \|\mathbf{b}-\mathbf{a}\|_s \\
\leq& K^{1/r}\|\mathbf{b}-\mathbf{a}\|_s,
\end{aligned} \tag{12}$$

where $r,s \geq 1$, $1/r + 1/s = 1$, the first inequality is due to Hölder's inequality, and the second inequality is due to the fact that $\nabla g(\mathbf{a}) = e^{\mathbf{a}}/\mathbf{1}'e^{\mathbf{a}}$, which implies that each element of $\nabla g(\mathbf{a})$ is smaller than 1. Based on (12) we have:

$$\begin{aligned}
&|\log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}_0}| \leq K^{1/r}\|\mathbf{B}'(\mathbf{x}-\mathbf{x}_0)\|_s \\
\leq& K^{1/r}\|\mathbf{B}'\|_s\|\mathbf{x}-\mathbf{x}_0\|_s,
\end{aligned} \tag{13}$$

where $r,s \geq 1$, $1/r + 1/s = 1$, and the last inequality is due to the definition of the matrix norm. For the second term of (11), we have,

$$|\mathbf{y}'\mathbf{B}'\mathbf{x} - \mathbf{y}'\mathbf{B}'\mathbf{x}_0| \leq \|\mathbf{y}\|_r\|\mathbf{B}'(\mathbf{x}-\mathbf{x}_0)\|_s \leq \|\mathbf{B}'\|_s\|\mathbf{x}-\mathbf{x}_0\|_s, \tag{14}$$

where the first inequality is due to Hölder's inequality, and the second inequality is due to the definition of the matrix norm and the fact that $\|\mathbf{y}\|_r = 1$. Combining (13) and (14), we have:

$$\begin{aligned}
&|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{x}_0,\mathbf{y})| \\
\leq& K^{1/r}\|\mathbf{B}'\|_s\|\mathbf{x}-\mathbf{x}_0\|_s + \|\mathbf{B}'\|_s\|\mathbf{x}-\mathbf{x}_0\|_s.
\end{aligned}$$

Under Assumptions A and B, by setting $\mathbf{x}_0 = \mathbf{0}$, we obtain that,

$$|h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) - h_{\mathbf{B}}(\mathbf{0},\mathbf{y})| \leq K^{1/r}\bar{C}R + \bar{C}R.$$

By noting that $h_{\mathbf{B}}(\mathbf{0},\mathbf{y}) = \log K$, we conclude: $h_{\mathbf{B}}(\mathbf{x},\mathbf{y}) \leq \log K + \bar{C}R(1+K^{1/r})$.

With the above results, the idea is to bound the expected loss using the empirical *Rademacher complexity* $\mathcal{R}_N(\cdot)$ of the class of loss functions: $\mathcal{H} = \{(\mathbf{x},\mathbf{y}) \to h_{\mathbf{B}}(\mathbf{x},\mathbf{y})\}$, denoted by $\mathcal{R}_N(\mathcal{H})$. Using Lemma 4.3.2 of [2] and the upper bound on the loss function, we arrive at the following result.

Lemma 1.3. *Under Assumptions A and B,*

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{2(\log K + \bar{C}R(1+K^{1/r}))}{\sqrt{N}}.$$

Using the Rademacher complexity of the class of loss functions, the out-of-sample prediction bias in Theorem 3.2 can be bounded by applying Theorem 8 in [3]. ☐

2. EXPERIMENTAL SETTINGS

We run the experiments on local GPU workstations with 4 NVIDIA RTX A6000 (48GB VRAM) and 2 NVIDIA Titan RTX (24GB VRAM) GPUs. The experiment for one epoch of DRO-MLR training on MNIST took only a few seconds while on CIFAR-10 it took about 0.05 GPU hours. Our ViT models were constructed under Huggingface Transformers v4.5.1 [4].

3. OMITTED EXPERIMENTAL RESULTS

We also implement DRO-MLR to Convolutional Neural Network (CNN) models. For a CNN image classifier, DRO-MLR is applied only to the last layer. We use a 10-layer Residual Network (ResNet) [5] on MNIST, and a 18-layer ResNet on CIFAR-10. The performance improvement of DRO-MLR shown in Fig. 1 is less significant compared to that in the ViT models, due to the fact that we only apply DRO-MLR to the last layer of CNN, while for ViT, DRO-MLR is applied to a larger set of layers.

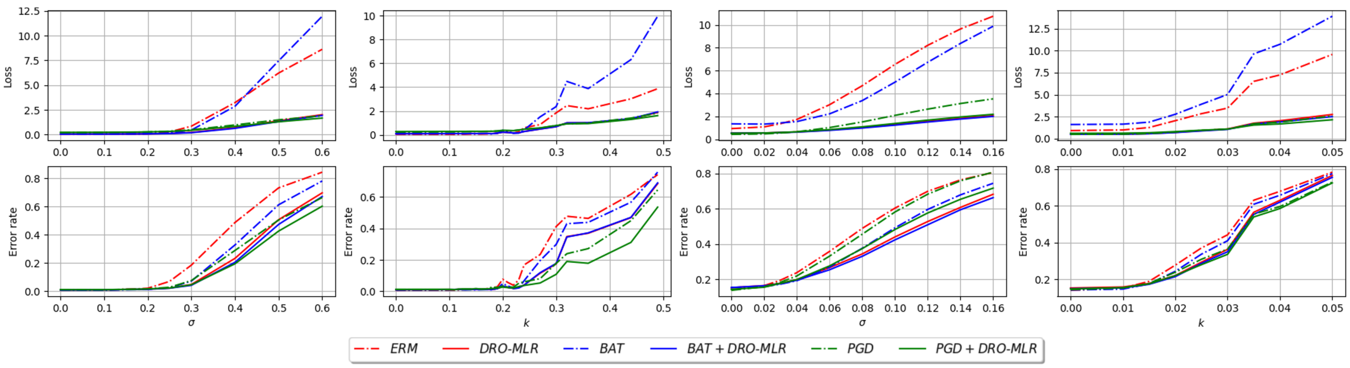
Finally, we briefly analyze the effect of applying DRO-MLR to different layers of ViT. In Fig. 2, DRO-MLR is applied separately to the final linear layer B, the initial patch projection layer P or the QKV-mapping layer in one of the self-attention layers. Compared with ERM, not all layers bring a significant performance improvement. The overall performance boost when all layers are re-trained with DRO-MLR can be largely credited to the B layer (the final linear layer). When DRO-MLR is applied only to the B layer, the loss is reduced by up to 87.0%, and the error rate is reduced by up to 67.6%, showing that re-training only the last linear layer using DRO-MLR is a fast and reliable way to improve the robustness of existing methods (as we did in CNN).

4. REFERENCES

[1] Peyman Mohajerin Esfahani and Daniel Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.

[2] Ruidi Chen and Ioannis Ch. Paschalidis, "Distributionally robust learning," *Foundations and Trends® in Optimization*, vol. 4, no. 1-2, pp. 1–243, 2020.

[3] Peter L Bartlett and Shahar Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.



(a) MNIST, $\mathcal{T}$ under WGN (b) MNIST, $\mathcal{T}$ under UAP (c) CIFAR-10, $\mathcal{T}$ under WGN (d) CIFAR-10, $\mathcal{T}$ under UAP

Fig. 1: Out-of-sample classification error and log-loss of different methods using CNN.

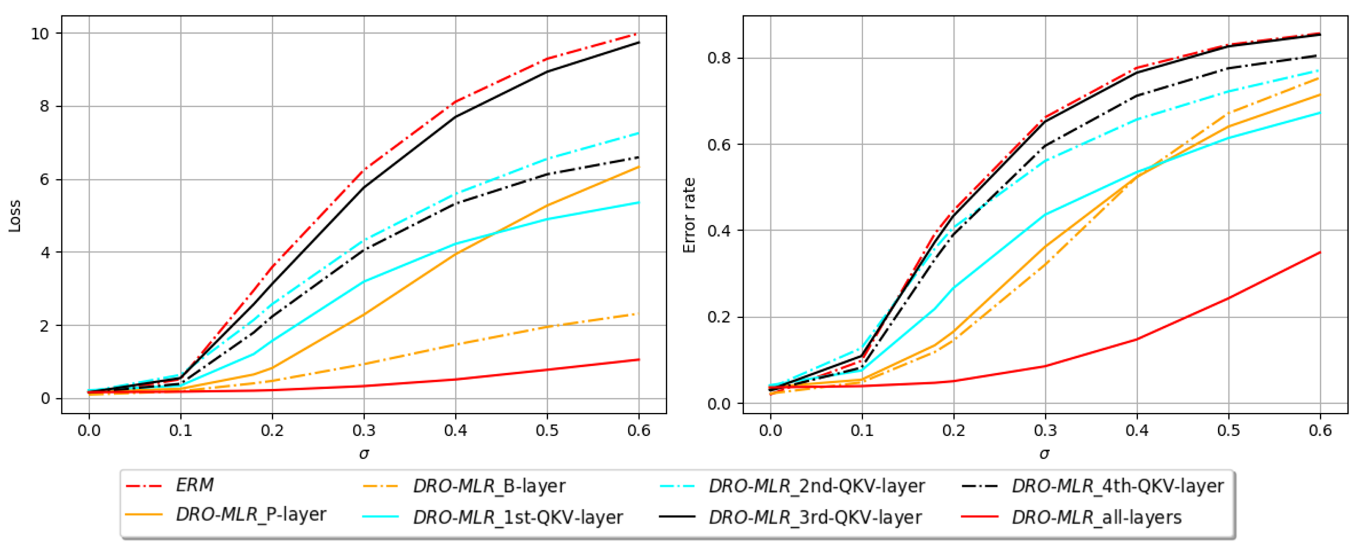


Fig. 2: Performance of applying DRO-MLR to different layers of ViT on MNIST under WGN.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.