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

ABSTRACT

We develop a *Distributionally Robust Optimization (DRO)* formulation for *Multiclass Logistic Regression (MLR)*, which could tolerate data contaminated by outliers. The DRO framework uses a probabilistic ambiguity set defined as a ball of distributions that are close to the empirical distribution of the training set in the sense of the Wasserstein metric. We relax the DRO formulation into a regularized learning problem whose regularizer is a norm of the coefficient matrix. We establish out-of-sample performance guarantees for the solutions to our model, offering insights on the role of the regularizer in controlling the prediction error. We apply the proposed method in rendering deep *Vision Transformer (ViT)*-based [1] image classifiers robust to random and adversarial attacks. Specifically, using the MNIST and CIFAR-10 datasets, we demonstrate reductions in test error rate by up to 83.5% and loss by up to 91.3% compared with baseline methods, by adopting a novel random training method.

Index Terms— Distributionally Robust Optimization, Multi-class Classification, Deep Learning.

1. INTRODUCTION

We consider the robust multi-class classification problem under the framework of *Distributionally Robust Optimization (DRO)*, where the ambiguity set is defined via the Wasserstein metric [2, 3]. Robust optimization has been widely used for inducing robustness to classification models [4, 5]. DRO, which minimizes the worst-case loss over a probabilistic ambiguity set, has received an increasing attention in recent years. The ambiguity set in DRO can be defined through moment constraints [6], or as a ball of distributions using some probabilistic distance function such as the Wasserstein distance [3]. The Wasserstein DRO model has been extensively studied in the machine learning community [7, 8, 9, 10, 11, 12, 13].

Most of the works on distributionally robust classification have focused on the binary setting [10]. In this paper, we extend the DRO framework to the multi-class setting by exploring *Multiclass Logistic Regression (MLR)* and deriving the corresponding robust model. We obtain a novel *matrix norm* regularizer for MLR through reformulating the DRO problem; thus, establishing a connection between robustness and regularization, and enabling a primal-dual interpretation for the data-regularizer relationship. Note that the link between robustness and regularization has been established in the 1-dimensional output setting, see, e.g., [14, 4, 15, 16, 17] for deterministic disturbances, and [10, 7, 8, 11, 12] for disturbances within a Wasserstein set. However, none of these works studied the robust problem with multi-dimensional outputs.

To the best of our knowledge, we are the first to study the robust multi-class classification problem from the standpoint of Wasserstein distributional robustness. Our problem is closely related to [10, 7] where the Wasserstein DRO problem with a 1-dimensional output was studied. The key differences lie in that: (i) [10, 7] only considered a scalar output y. We make the non-trivial extension to a multi-class scenario, which is very different and our key and novel contribution. Specifically, in this work we are regularizing a coefficient matrix $\mathbf{B}$ and the correlation structure embedded in the responses should be reflected in the regularizer which is derived as the dual norm of the matrix; (ii) the out-of-sample result in Theorem 3.1 is tailored to the classification setting, which is different from the regression setting in [7], and from [10] which bounds the out-of-sample risk by the optimal worst-case risk; and (iii) we demonstrate a computationally efficient random-layer training mechanism for applying DRO in *Vision Transformer (ViT)*-based [1] image classifiers, which adds to the accessibility and appeal of this work to practitioners.

The rest of the paper is organized as follows. In Section 2, we develop the Wasserstein DRO formulation for MLR. Section 3 establishes the out-of-sample performance guarantees for the DRO solution. Numerical experimental results with deep *Vision Transformer (ViT)*-based [1] image classifiers are presented in Section 4. We conclude the paper in Section 5.

Notational convention. We use boldfaced lowercase letters to denote vectors, ordinary lowercase letters to denote scalars, boldfaced uppercase letters to denote matrices, and calligraphic capital letters to denote sets. All vectors are column vectors. For space saving reasons, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$

is the dimension of $\mathbf{x}$. We use prime to denote the transpose, $\|\cdot\|_p$ for the ℓ_p norm with $p \geq 1$, and $\|\mathbf{x}\|_p^{\mathbf{W}}$ for the $\mathbf{W}$-weighted ℓ_p norm defined as $\|\mathbf{x}\|_p^{\mathbf{W}} \triangleq \|\mathbf{W}^{1/2}\mathbf{x}\|_p$, with a positive definite matrix $\mathbf{W}$. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we use $\|\mathbf{A}\|_p$ to denote its induced ℓ_p norm that is defined as $\|\mathbf{A}\|_p \triangleq \sup_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{A}\mathbf{x}\|_p / \|\mathbf{x}\|_p$. Finally, $\mathbf{1}$ denotes the vector of ones, $\mathbf{0}$ the vector of zeros, and $\mathbf{e}_k$ the k-th unit vector.

2. PROBLEM FORMULATION

Suppose there are K classes, and we are given a predictor $\mathbf{x} \in \mathbb{R}^p$. Our goal is to predict its class label, denoted by a K-dim vector $\mathbf{y} \in \{0,1\}^K$, where $\mathbf{y} = (y_1, \ldots, y_K)$, $\sum_k y_k = 1$, and $y_k = 1$ if and only if $\mathbf{x}$ belongs to class k, in which case $\mathbf{y} = \mathbf{e}_k$. The conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is modeled as $p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^K p_i^{y_i}$, where $p_i = e^{\mathbf{w}_i'\mathbf{x}} / \sum_{k=1}^K e^{\mathbf{w}_k'\mathbf{x}}$, and $\mathbf{w}_i, i \in [\![K]\!]$, are the coefficient vectors to be estimated. The log-likelihood can be expressed as:

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^K y_i \log(p_i) = \mathbf{y}'\mathbf{B}'\mathbf{x} - \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}},$$

where $\mathbf{B} \triangleq [\mathbf{w}_1 \cdots \mathbf{w}_K]$ and the exponential operator is applied element-wise. The log-loss is defined as $h_{\mathbf{B}}(\mathbf{x}, \mathbf{y}) \triangleq \log \mathbf{1}'e^{\mathbf{B}'\mathbf{x}} - \mathbf{y}'\mathbf{B}'\mathbf{x}$. The Wasserstein DRO formulation for MLR minimizes over $\mathbf{B}$ the worst-case expected loss:

$$\inf_{\mathbf{B}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}\big[h_{\mathbf{B}}(\mathbf{x}, \mathbf{y})\big], \qquad (1)$$

where $\mathbb{E}^{\mathbb{Q}}$ denotes expectation under a distribution $\mathbb{Q}$ of the data $(\mathbf{x}, \mathbf{y})$, with $\mathbb{Q}$ belonging to a set Ω:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_1(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\}, \qquad (2)$$

where $\mathcal{Z}$ is the set of possible values for $(\mathbf{x}, \mathbf{y})$, i.e., $\mathcal{Z} = \mathbb{R}^p \times \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$; $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$; ϵ is a pre-specified positive constant; and $\hat{\mathbb{P}}_N$ is the empirical distribution that assigns equal probability to each observed sample $(\mathbf{x}_i, \mathbf{y}_i), i \in [\![N]\!]$. $W_1(\mathbb{Q}, \hat{\mathbb{P}}_N)$ is the order-1 Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$:

$$W_1(\mathbb{Q}, \hat{\mathbb{P}}_N) \triangleq \min_{\Pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} \left\{ \int_{\mathcal{Z} \times \mathcal{Z}} l(\mathbf{z}_1, \mathbf{z}_2) \, \Pi(\mathrm{d}\mathbf{z}_1, \mathrm{d}\mathbf{z}_2) \right\}, \qquad (3)$$

where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i), i = 1, 2$, Π is the joint distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$, respectively, and $l(\cdot, \cdot)$ is a distance metric on the data space that measures the cost of transporting the probability mass and is defined as:

$$l(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_r + M\|\mathbf{y}_1 - \mathbf{y}_2\|_t, \qquad (4)$$

with a positive constant M.

Theorem 2.1 derives an equivalent reformulation of (1) by exploiting the dual problem of the inner supremum in (1). The proof can be found in the Supplement[1].

---

[1] Supplementary material link: https://github.com/noc-lab/dro_mlr

Theorem 2.1. *Suppose we observe N realizations of the data, denoted by $(\mathbf{x}_i, \mathbf{y}_i), i \in [\![N]\!]$. When the Wasserstein metric is induced by (4), as $M \to \infty$, the DRO problem (1) can be reformulated as:*

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^N h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon 2^{1/s} \|\mathbf{B}\|_s, \qquad (5)$$

where $r, s \geq 1$, and $1/r + 1/s = 1$. We call (5) the DRO-MLR formulation.

Remark: a weighted norm space. When the feature space is equipped with a weighted norm, e.g.,

$$l(\mathbf{z}_1, \mathbf{z}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_r^{\mathbf{W}} + M\|\mathbf{y}_1 - \mathbf{y}_2\|_t, \qquad (6)$$

where $\|\mathbf{x}\|_r^{\mathbf{W}} \triangleq \|\mathbf{W}^{1/2}\mathbf{x}\|_r$, with $\mathbf{W}$ a positive definite matrix, the corresponding DRO-MLR formulation can be written as:

$$\inf_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^N h_{\mathbf{B}}(\mathbf{x}_i, \mathbf{y}_i) + \epsilon 2^{1/s} \|\mathbf{W}^{-1/2}\mathbf{B}\|_s, \qquad (7)$$

where $r, s \geq 1, 1/r + 1/s = 1$. This is due to the fact that the dual norm of $\|\cdot\|_r^{\mathbf{W}}$ is simply $\|\cdot\|_s^{\mathbf{W}^{-1}}$.

3. OUT-OF-SAMPLE PERFORMANCE

In this section we establish out-of-sample performance guarantees for the DRO-MLR solution, i.e., given a new test sample, we bound the expected log-loss of our prediction. The resulting bounds shed light on the role of the regularizer in inducing a low prediction error.

We want to measure the out-of-sample performance in terms of the empirical loss, which is typically used in *Empirical Risk Minimization (ERM)*, so that we can illustrate the advantage of DRO-MLR compared to ERM. By bounding the *Rademacher complexity* of the class of loss functions, Theorem 3.1 bounds the expected log-loss by the empirical loss plus additional terms that are inversely proportional to $\sqrt{N}$.

Assumption A. *For any $\mathbf{x}$: $\|\mathbf{x}\|_s \leq R$ almost surely, for some scalar R.*

Assumption B. *For any feasible solution $\mathbf{B}$ to (5): $\|\mathbf{B}'\|_s \leq \bar{C}$, for some scalar $\bar{C}$.*

With standardized predictors, R in Assumption A can be assumed to be small. The form of the constraint in Assumptions B is consistent with the form of the regularizers in DRO-MLR. We will see later that the bound $\bar{C}$ controls the out-of-sample log-loss of the solution to DRO-MLR, which validates the role of the regularizer in improving the out-of-sample performance.

Theorem 3.1. *Suppose the solution to the DRO-MLR formulation (5) is $\hat{\mathbf{B}}_N$. Under Assumptions A and B, for any*

$0 < \alpha < 1$, *with probability at least* $1 - \alpha$ *with respect to the sampling, and with* $\mathbb{P}^*$ *the true measure,*

$$\mathbb{E}^{\mathbb{P}^*}[h_{\hat{\mathbf{B}}_N}(\mathbf{x}, \mathbf{y})] \leq \mathbb{E}^{\hat{\mathbb{P}}_N}[h_{\hat{\mathbf{B}}_N}(\mathbf{x}, \mathbf{y})] + \frac{2\left(\log K + \bar{C}R(1 + K^{1/r})\right)}{\sqrt{N}}$$

$$+ \left(\log K + \bar{C}R(1 + K^{1/r})\right)\sqrt{\frac{8\log(\frac{2}{\alpha})}{N}} \ .$$

Theorem 3.1 says that the out-of-sample generalization (test) error of the DRO-MLR solution is bounded by the average training error plus a bias term of order $1/\sqrt{N}$. The bound in Theorem 3.1 demonstrates the validity of DRO-MLR in leading to a good out-of-sample performance.

4. DRO-MLR IN DEEP IMAGE CLASSIFIERS

We apply DRO-MLR to deep ViT based image classifiers, and provide an efficient mechanism for inducing robustness to random and adversarial attacks in deep neural networks through applying DRO to a random layer at each epoch, which is computationally efficient and can be generalized to any deep learning-based approaches. We adopt a metric learning approach to estimate an appropriate norm for defining the Wasserstein metric. Our method is compared with ERM and other adversarial training methods to demonstrate the effectiveness of our model in inducing a smaller generalization error and test error rate.

We split the dataset into a training set $\mathcal{D}$, a validation set $\mathcal{V}$ where hyper-parameter tuning (e.g., the regularization coefficient ϵ) is performed, and a test set $\mathcal{T}$. A perturbed image $\tilde{\mathbf{x}}$ is generated as $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is generated by either a random attack such as *White Gaussian Noise (WGN)*, or an adversarial attack such as *Universal Adversarial Perturbation (UAP)* [18]. The WGN attack perturbs each pixel by a Gaussian noise with zero mean and standard deviation σ, while the UAP attack is generated using $10,000$ images from the training set $\mathcal{D}$, with $\|\boldsymbol{\delta}\|_\infty \leq k$, and k a pre-specified perturbation parameter. UAP has been shown to misclassify new images with high probability irrespective of the network architecture.

We consider three baseline methods: (1) *Empirical Risk Minimization (ERM)*, which simply minimizes the sample averaged log-loss; (2) *Brute-force Adversarial Training (BAT)*, which adds adversarial samples into the training set $\mathcal{D}$ and performs ERM; and (3) *Projected Gradient Descent (PGD)* [19], which iteratively finds an optimal perturbation vector by using the gradient of the trained model from the previous iteration, and then updates the model using the perturbed samples. BAT is widely used in practice since it can be easily implemented, and PGD is known to be a strong defensive approach against many types of adversarial attacks. We compare these methods with their variations where DRO-MLR is applied, in terms of the log-loss and classification error on the test set $\mathcal{T}$.

We use $r = 2$ to define the distance metric (4). It is worth noting that when the DRO-MLR model is applied to a layer l,

Table 1: Percentage improvement (Mean and Standard deviation) of DRO-MLR over three baseline methods using ViT at maximum attack strength under WGN and UAP attacks on the test set. Percentage improvement is defined as the ratio of the change in the metric (error rate or loss) over the base metric value.

		MNIST				CIFAR-10			
		WGN		UAP		WGN		UAP	
		Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
ERM	Loss	89.4%	0.1%	69.5%	1.5%	32.2%	0.8%	79.1%	0.2%
	Error rate	58.1%	1.3%	19.3%	0.8%	37.7%	0.5%	83.5%	0.3%
BAT	Loss	83.6%	0.7%	79.5%	0.6%	64.4%	0.9%	78.3%	1.7%
	Error rate	42.8%	2.5%	19.8%	2.5%	15.8%	1.3%	76.3%	1.9%
PGD	Loss	91.3%	0.2%	79.9%	1.5%	59.9%	0.9%	53.4%	1.3%
	Error rate	61.0%	2.3%	28.8%	7.3%	43.6%	1.7%	35.7%	1.5%

whose input is denoted as $\phi_l(\mathbf{x})$ which takes into account the non-linear transformation of the raw image $\mathbf{x}$ resulted from all layers before l, we need to estimate a proper distance metric in the transformed space to account for the distributional shift resulted from ϕ_l. We adopt a weighted norm metric as defined in (6) to approximate the effect of ϕ_l, and estimate the weight matrix $\mathbf{W}$ by solving the following metric learning problem on the training set $\mathcal{D}$:

$$\min_{\mathbf{W} \succcurlyeq 0} \quad \sum_{\mathbf{x}_i \in \mathcal{D}} \|\mathbf{W}^{1/2}(\phi_l(\tilde{\mathbf{x}}_i) - \phi_l(\mathbf{x}_i))\|_2^2$$
$$\text{s.t.} \quad \frac{1}{|\mathcal{S}|} \sum_{(i,j) \in \mathcal{S}} \|\mathbf{W}^{1/2}(\phi_l(\tilde{\mathbf{x}}_i) - \phi_l(\tilde{\mathbf{x}}_j))\|_2^2 \geq c, \quad (8)$$
$$\frac{1}{|\mathcal{S}|} \sum_{(i,j) \in \mathcal{S}} \|\mathbf{W}^{1/2}(\phi_l(\mathbf{x}_i) - \phi_l(\mathbf{x}_j))\|_2^2 \geq c,$$

where $\tilde{\mathbf{x}}_i$ is the perturbed version of $\mathbf{x}_i$, $\mathcal{S} \triangleq \{(i,j)|\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}, \mathbf{y}_i \neq \mathbf{y}_j\}$, $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$, and c is a fixed parameter. Note that for ViT where there exist multiple patches, we simply add up the squared norm in the objective and constraints of (8) over patches. Problem (8) enforces that the distances between similar samples (evaluated in the transformed space $\phi_l(\mathbf{x})$) are being minimized while the distances between dissimilar samples are sufficiently far away. (8) is a semidefinite programming problem (SDP) which can be solved using SDPT3 solver [20]. In order to speed up the training, in each epoch, we only apply DRO-MLR to one random layer while keeping all other layers frozen, which has a similar flavor to the fine-tuning strategy of [21] that modifies the parameters of an existing network to train for a new task while preserving representations learned from the original task. Our novel random-layer training approach speeds up the training and bears a similar spirit to random dropout where efficiency and robustness are balanced through randomness.

We demonstrate the results on MNIST [22] and CIFAR-10 [23], which contain 50k/10k/10k and 40k/10k/10k training/validation/test samples, respectively. We apply a ViT with 4 attention layers on MNIST, and on CIFAR-10, we fine-tuned a ViT-B/16 model [1]. We used the learning rate of 1×10^{-3} for MNIST and 1×10^{-4} for CIFAR-10, while no weight decay
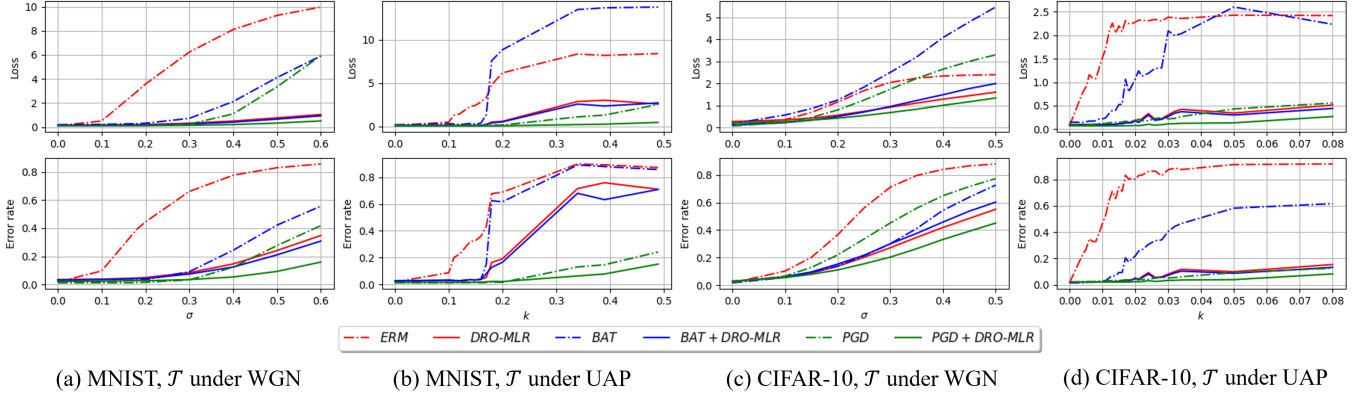
Fig. 1: Out-of-sample classification error and log-loss of different methods using ViT.

was applied. The results are shown in Fig. 1, where we plot the average log-loss $h_\mathcal{T}$ and classification error $e_\mathcal{T}$ on the test set $\mathcal{T}$ for various methods under different attacks, as the perturbation strength σ or k varies. We see that when DRO-MLR is combined with various adversarial training methods, both the loss and error rate are significantly reduced, with PGD+DRO-MLR outperforming the rest. The performance gap becomes more prominent as the perturbation strength increases. The results suggest that DRO-MLR can be potentially combined with any existing adversarial training method on any neural network structure to further improve its performance.

Table 1 summarizes the reduction in the loss and error rate at maximum σ and k. On MNIST, when ERM / BAT is combined with DRO-MLR, the test error rate is reduced by up to 58% / 43% and log-loss is reduced by up to 89% / 84%, respectively. On CIFAR-10, the reductions w.r.t. ERM / BAT are 84% / 76% for error rate, and 79% / 78% for log-loss. Note that PGD remains a powerful defense that it can sometimes outperform the vanilla DRO-MLR model under the adversarial attack. Nevertheless, when combined with DRO-MLR, PGD can be further improved by up to 61% / 91% for error rate / log-loss on MNIST, and up to 44% / 60% on CIFAR-10. On the other hand, Fig. 1 indicates that DRO-MLR has a comparable performance to ERM under very small perturbations, implying that DRO-MLR is able to induce robustness to perturbations without compromising the accuracy on unperturbed data.

We want to emphasize that during the training, both DRO-MLR and BAT make use of the adversarial samples. The fact that DRO-MLR outperforms BAT suggests that DRO-MLR uses a more efficient way to leverage the information in the adversarial samples (to learn the $\mathbf{W}$ matrix) without expanding the training set, which could also explain the significant improvement that DRO-MLR brings to PGD, demonstrating the potential of DRO-MLR in being combined with other adversarial training methods to make further improvement.

Finally, visualizing the attention maps in ViT can also help to understand the advantages of DRO-MLR. In Fig. 2 (a) and
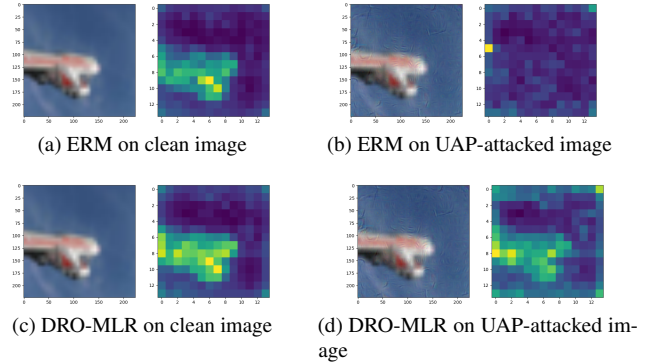
Fig. 2: Visualization of the attention maps from the ViT.

(c), both the ERM and DRO-MLR models can capture the most important patches in the clean images through the self-attention mechanism. Nevertheless, when the image is slightly perturbed by the UAP attack in Fig. 2 (b), self-attention under ERM loses its strength and fails to capture the important area, while DRO-MLR can still clearly find the position of the airplane in Fig. 2 (d). This serves as a strong evidence that DRO-MLR can help ViT preserve reliable self-attention.

5. CONCLUSION

We proposed a novel distributionally robust framework under the Wasserstein metric for *Multiclass Logistic Regression (MLR)*, and reformulated the min-max formulation to a regularized empirical loss minimization problem, establishing a connection between robustness and regularization in the multivariate setting. We provide both theoretical results on the performance of our estimator, and empirical evidence on deep ViT-based image classifiers, showing that our DRO-MLR model reduces the baseline test error by up to 83.5% and loss by up to 91.3% under random and adversarial attacks.

6. REFERENCES

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] Rui Gao and Anton J Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[3] 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.

[4] Laurent El Ghaoui, Gert René Georges Lanckriet, and Georges Natsoulis, "Robust classification with interval data," 2003.

[5] Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo, "Robust classification," *INFORMS Journal on Optimization*, vol. 1, no. 1, pp. 2–34, 2018.

[6] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014.

[7] Ruidi Chen and Ioannis Ch Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 517–564, 2018.

[8] Jose Blanchet, Peter W Glynn, Jun Yan, and Zhengqing Zhou, "Multivariate distributionally robust convex regression under absolute error loss," *arXiv preprint arXiv:1905.12231*, 2019.

[9] Aman Sinha, Hongseok Namkoong, and John Duchi, "Certifiable distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.

[10] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn, "Distributionally robust logistic regression," in *Advances in Neural Information Processing Systems*, 2015, pp. 1576–1584.

[11] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani, "Regularization via mass transportation," *arXiv preprint arXiv:1710.10016*, 2017.

[12] Rui Gao, Xi Chen, and Anton J Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," *arXiv preprint arXiv:1712.06050*, 2017.

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

[14] Laurent El Ghaoui and Hervé Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal on Matrix Analysis and Applications*, vol. 18, no. 4, pp. 1035–1064, 1997.

[15] Huan Xu, Constantine Caramanis, and Shie Mannor, "Robustness and regularization of support vector machines," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1485–1510, 2009.

[16] Huan Xu, Constantine Caramanis, and Shie Mannor, "Robust regression and LASSO," in *Advances in Neural Information Processing Systems*, 2009, pp. 1801–1808.

[17] Dimitris Bertsimas and Martin S Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 2017.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[20] Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü, "SDPT3 — a MATLAB software package for semidefinite programming, version 1.3," *Optimization methods and software*, vol. 11, no. 1-4, pp. 545–581, 1999.

[21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[22] Yann LeCun, "The MNIST database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[23] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.