ROBUST GENERALIZATION AGAINST CORRUPTIONS VIA WORST-CASE SHARPNESS MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust generalization aims to deal with the most challenging data distributions which are rarely presented in training set and contain severe noise corruptions. Common solutions such as distributionally robust optimization (DRO) focus on the worst-case empirical risk to ensure low training error on the uncommon noisy distributions. However, due to the over-parameterized model being optimized on scarce worst-case data, DRO fails to produce a smooth loss landscape, thus struggling on generalizing well to the test set. Therefore, instead of focusing on the worst-case risk minimization, we propose SharpDRO by penalizing the sharpness of the worst-case distribution, which measures the loss changes around the neighbor of learning parameters. Through worst-case sharpness minimization, the proposed method successfully produces a flat loss curve on the corrupted distributions, thus achieving robust generalization. Moreover, by considering whether the distribution annotation is available, we apply SharpDRO to two problem settings and design a worst-case selection process for robust generalization. Through simulating real-world noisy distributions using CIFAR10/100 and ImageNet30 datasets, we show that SharpDRO exhibits strong generalization ability against severe corruptions and exceeds well-known baseline methods with large performance gains.

1 INTRODUCTION

Learning against noise corruption has been a vital challenge in the practical deployment of machine learning, as the learning models are much more fragile to subtle perturbations than human perception systems (Goodfellow et al., 2015; Hendrycks & Gimpel, 2017). During the training process, the encountered corruptions are essentially perceived as distribution shift, which would significantly hinder the learning results (Liang et al., 2018; Long et al., 2015; Tzeng et al., 2017). Therefore, to mitigate the performance degradation, learning a robust model that generalizes well to corrupted data distributions has drawn lots of attention (Arjovsky et al., 2019; Sagawa et al., 2020).

In the real world, noise corruptions often come with different levels of severity. As a result, such a variety of severity would form multiple data distributions which impose varied negative impacts on our learning model (Hendrycks & Dietterich, 2019). Specifically, we assume the encountered corruption $\mathscr{E}$ is a composition of multiple noise unit u. Each noise unit is triggered by some discrete factors that appear with a certain probability during a given time interval. For example, a noise unit u is caused by one single data compression process which would happen during each platform changing, re-distribution, transmission, and so on. More compression is conducted, and severer corruption would be applied to the data. Therefore, the severity s of the corruption $\mathscr{E}$ can be reasonably modeled by Poisson distribution, *i.e.*, $s \sim P(s;\lambda) = \frac{e^{-\lambda}\lambda^s}{s!}$, which is illustrated in Figure 1. As a result, the real-world training set is not completely composed of clean data, but contains corrupted data with a smaller proportion as the severity goes stronger.

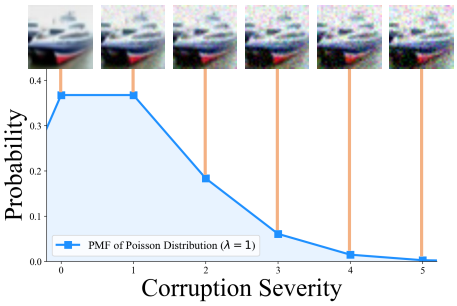

Figure 1: Illustration of real-world noisy distributions. We take Gaussian Noise for example.

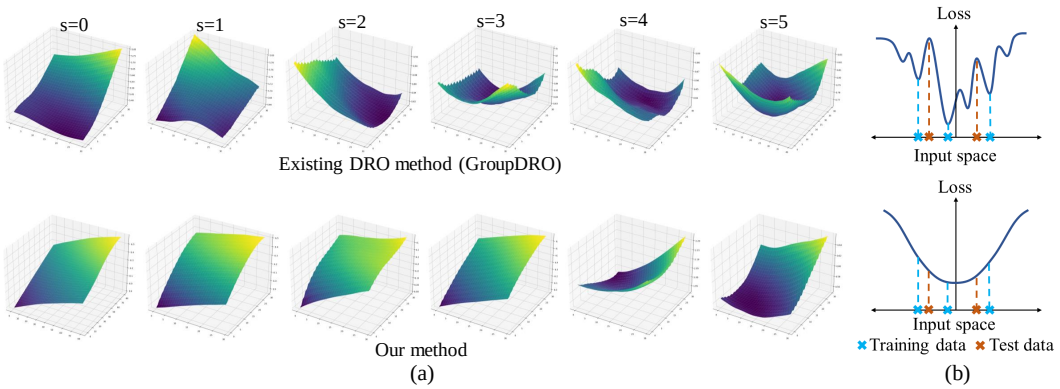

Figure 2: Illustration of our motivation. (a) Visualization of the loss surface of GroupDRO and the proposed SharpDRO. The columns from left to right stand for corrupted distributions with severity $s = 0$ to 5. (b) Illustration of why a sharp loss surface hinders generalization to test data.

Dealing with such a realistic problem by vanilla empirical risk minimization can achieve satisfactory averaged accuracy on the whole training set. However, due to the extremely limited number of severely-corrupted data, the learning model would produce large training errors on the corrupted distributions, further hindering the robust performance under challenging real-world situations. A popular approach to achieving low error on the scarce corrupted data is distributionally robust optimization (DRO) (Namkoong & Duchi, 2016; Sagawa et al., 2020; Zhai et al., 2021), which commonly optimizes the model parameter θ by the following objective:

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{L}(\theta; (x, y)) \right], \tag{1}$$

where $\mathcal{Q}$ denotes the uncertainty set that is utilized to estimate the possible test distribution. Intuitively, DRO assumes that $\mathcal{Q}$ consists of multiple sub-distributions, among which exists a worst-case distribution Q. By concentrating on the risk minimization of the worst-case distribution, DRO hopes to train a robust model which can deal with the potential distribution shift during the test phase. However, existing DRO methods usually leverage over-parameterized models to focus on a small portion of worst-case training data. Therefore, the worst-case data contaminated with severe corruption is highly possible to get stuck into sharp minima. As shown in the upper of Figure 2 (a), a stronger corruption severity would cause existing method to learn a sharper loss surface. Consequently, optimization via DRO fails to produce a flat loss landscape over the corrupted distributions, which further leads to a large generalization gap between training and test set (Keskar et al., 2017; Chaudhari et al., 2017).

To remedy this defect, in this paper, we propose SharpDRO method to focus on learning a flat loss landscape of the worst-case data, which can largely mitigate the training-test generalization gap problem of DRO. Specifically, we adopt the largest loss difference formed by applying weight perturbation (Foret et al., 2020; Wu et al., 2020) to measure the sharpness of the loss function. Intuitively, a sharp loss landscape is sensitive to noise and cannot generalize well on the test set. On the contrary, a flat loss landscape produces consistent loss values and is robust against perturbations (Figure 2 (b)). By minimizing the sharpness, we can effectively enhance the generalization performance. However, directly applying sharpness minimization on multiple distributions would yield poor results Cha et al. (2021), as the computed sharpness could be influenced by the largest data distribution, and thus cannot generalize well to small corrupted data. Therefore, we only focus on worst-case sharpness minimization. In this way, as the lower of Figure 2 (a) shows, SharpDRO successfully produces a flat loss surface, thus achieving robust generalization on the severely corrupted distributions.

In addition, identification of the worst-case distribution requires expensive annotations, which are not always practically feasible (Liu et al., 2021). In this paper, we apply SharpDRO to solve two problem settings: 1) *Distribution-aware robust generalization* which assumes that distribution indexes are accessible, and 2) *Distribution-agnostic robust generalization* where the distributions are no longer identifiable, making the worst-case data hard to find. Existing approaches such as Just Train Twice (JTT) require two-stage training which is rather inconvenient. To tackle this challenge, we propose a simple OOD detection (Hendrycks & Gimpel, 2017; Liang et al., 2018) process to detect the worst-case data, which can be further leveraged to enable worst-case sharpness minimization. Through constructing training sets according to the Poisson distributed noisy distribution using CIFAR10/100 and ImageNet30, we show that SharpDRO can achieve robust generalization results on both two problem settings, surpassing well-known baseline methods by a large margin.

To sum up, our contributions are threefold:

- We proposed a sharpness-based DRO method that overcomes the poor generalization performance on worst-case distribution in distributionally robust optimization.
- We apply the proposed SharpDRO method to both distribution-aware and distribution-agnostic settings, which brings a practical capability to our method. Moreover, we propose an OOD detection approach to select worst-case data to enable robust generalization.
- We form a real-world noisy training set that follows Poisson distribution, and conduct extensive analyses to show a strong generalization ability of SharpDRO as well as its superiority to compared baseline methods.

This paper is organized as follows: In Section 2, we first briefly introduce several well-known baseline methods and give details about the problem setting. Then we specify our worst-case sharpness minimization over two problem settings in Section 3. To validate the proposed method, we empirically support our SharpDRO by carefully conducting experimental analyses in Section 4. At last, we conclude this paper in Section 5.

2 Robust Generalization Methods against Distribution Shift

Due to the practical significance of robust generalization, various approaches have been proposed to deal with distribution shift. Here we briefly introduce three typical baseline methods, namely Invariant Risk Minimization, Risk Extrapolation, and GroupDRO.

Invariant Risk Minimization (IRM) (Arjovsky et al., 2019; Chang et al., 2020; Creager et al., 2021) aims to extract the invariant feature across different distributions (also denoted as environments). Specifically, the learning model is separated into a feature extractor G and a classifier C. IRM assumes an invariant model $C \circ G$ over various environments can be achieved if the classifier C constantly stays optimal. Then, the learning objective is formulated as:

$$\begin{aligned} &\min_{C^* \circ G} \left\{ \mathcal{L}_{\text{IRM}} := \sum_{e \in \mathcal{E}} \mathcal{L}^e(C^* \circ G) \right\} \\ &\text{subject to } C^* \in \arg\min_{G} \mathcal{L}^e(C \circ G), \text{for all } e \in \mathcal{E}, \end{aligned} \tag{2}$$

where C^* stands for the optimal classifier, and e denotes a specific environment from a given environmental set $\mathcal{E}$. By solving Eq. 2, the feature extractor G can successfully learn invariant information without being influenced by the distribution shift between different environments.

Risk Extrapolation (REx) (Krueger et al., 2021) targets at generalization to out-of-distribution (OOD) environments. Inspired by the discovery that penalizing the loss variance across distributions helps achieve improved performance on OOD generalization, REx proposes to optimizing via:

$$\min_{\theta \in \Theta} \left\{ \mathcal{L}_{\text{REx}} := \sum_{e \in \mathcal{E}} \mathcal{L}^e(\theta) + \beta Var(\mathcal{L}^e, ..., \mathcal{L}^m) \right\}, \tag{3}$$

where β controls the penalization level. Intuitively, REx seeks to achieve risk fairness among all m training environments, so as to increase the similarity of different learning tasks. As a result, the training model can capture the invariant information that helps generalize to unseen distributions.

GroupDRO (Sagawa et al., 2020; Hashimoto et al., 2018; Piratla et al., 2021) deal with the situation when the correlation between class label y and unknown attribute a differs in the training and test set. Such a difference is called spurious correlation which could seriously misguide the model prediction. As a solution, GroupDRO considers each combination of class and attribute as a group g. By conducting risk minimization though:

$$\min_{\theta \in \Theta} \left\{ \mathcal{L}_{\text{GroupDRO}} := \max_{g} \mathbb{E}_{(x,y) \sim P_g} \left[\mathcal{L}(\theta; (x, y)) \right] \right\}, \tag{4}$$

the worst-case group from distribution P_s which commonly holds spurious correlation is emphasized, thus breaking the spurious correlation.

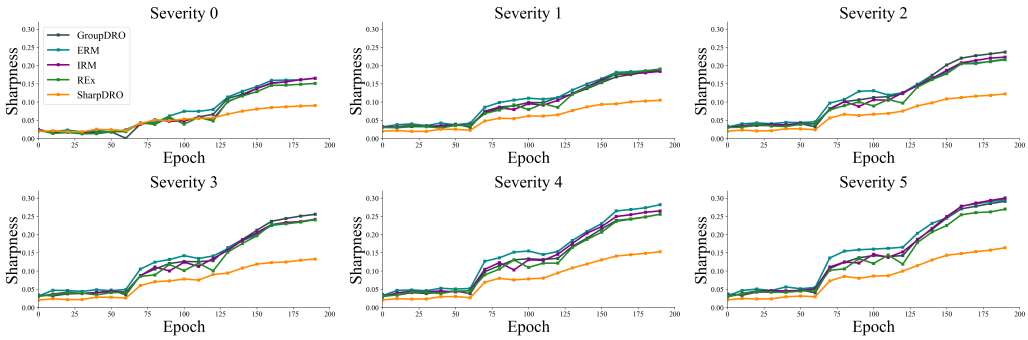

Figure 3: Sharpness during networking training on clean ($s = 0$) and corrupted distributions ($s = 1$ to 5).

Discussion: IRM and REx both focus on learning invariant knowledge across various environments. However, when the training set contains extremely imbalanced noisy distributions, as shown in Figure 1, the invariant learning results would be greatly misled by the most dominating distribution. Thus, the extracted invariant feature would be questionable for generalization against distribution shift. Although emphasizing the risk minimization of worst-case data via GroupDRO can alleviate the imbalance problem, its generalization performance is still sub-optimal when facing novel test data. However, SharpDRO can not only focus on the uncommon corrupted data but also effectively improve the generalization performance on the test set by leveraging worst-case sharpness minimization. In the next section, we elaborate on the methodology of SharpDRO.

3 METHODOLOGY

In robust generalization problems, we are given a training set $\mathcal{D}_{\text{train}}$ containing n image examples, each example $x \in \mathcal{X}$ is given a class label $y \in \mathcal{Y} = \{1, 2, ..., c\}$. Moreover, the training set is corrupted by a certain type of noise whose severity s follows a Poisson distribution $P(s; \lambda)$. Here we assume $\lambda = 1$ which indicates that the mean number of the noise unit u that occurred during a time interval is 1. Therefore, the distribution P of the whole training set is composed of S sub-distributions $P_s, s \in \{1, 2, ..., S\}$ with varied levels of corruption. Our goal is to learn a robust model $\theta \in \Theta$ that can achieve good generalization performance on challenging data distributions P_s with large severity.

In general, the learning objective of our SharpDRO can be formulated as:

$$\min_{\theta} \left\{ \mathcal{L}_{\text{SharpDRO}} := \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{L}(\theta; (x, y)) \right] + \mathbb{E}_{(x,y)\sim Q} \left[\mathcal{R}(\theta; (x, y)) \right] \right\}, \tag{5}$$

where the first term denotes the risk minimization using loss function $\mathcal{L}$, meanwhile a worst-case distribution Q is selected based on the model prediction. The second term indicates the sharpness minimization which aims to maximally improve the generalization performance on the worst-case distribution Q. Specifically, as shown in Figure 3, the sharpness gradually increases as the corruption severity enlarges. Therefore, to accomplish robust generalization, we are motivated to emphasize the worst-case distribution. As a result, we can produce much smaller sharpness compared to other methods, especially on the severely corrupted distributions.

In the following content, we first introduce sharpness on worst-case data for robust generalization. Then, we demonstrate our worst-case data selection on two problem settings. Finally, we give details on the practical implementation of SharpDRO.

3.1 SHARPNESS FOR ROBUST GENERALIZATION

The main challenge of robust generalization is that the training distribution is extremely imbalanced, as shown in Figure 1. The learning performance on the abundant clean data is quite satisfactory, but robustness regarding the corrupted distribution is highly limited, owing to the severe disturbance of corruption as well as the insufficiency of noisy data. To enhance the robust generalization performance, we leverage sharpness to fully exploit the worst-case data. Specifically, sharpness (Foret et al., 2020; Wu et al., 2020; Zheng et al., 2021) is measured by the largest loss change when model weight θ is perturbed with ϵ, which is formulated as

$$\mathcal{R} := \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon; (x, y)) - \mathcal{L}(\theta; (x, y)), \tag{6}$$

where ρ is a scale parameter to control the perturbation magnitude. By supposing the weight perturbation is small enough, we can have:

$$\mathcal{L}(\theta + \epsilon) - \mathcal{L}(\theta) \approx \nabla \mathcal{L}(\theta)\epsilon. \tag{7}$$

Further, we hope to obtain the largest loss difference to find the optimal weight perturbation ϵ^*, which can be computed as:

$$\epsilon^* := \arg\max_{\|\epsilon\| \leq \rho} \nabla \mathcal{L}(\theta)\epsilon. \tag{8}$$

By following dual norm problem, the optimal ϵ^* can be solved as $\rho\,\mathrm{sign}(\nabla\mathcal{L}(\theta))$ (Foret et al., 2020), which is essentially the ∞-norm of the gradient $\nabla\mathcal{L}$ multiplied with a scale parameter ρ. To this end, our sharpness minimization can be formulated as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim Q}\mathcal{R} := \mathcal{L}(\theta + \rho\,\mathrm{sign}(\nabla\mathcal{L}(\theta;(x,y)))) - \mathcal{L}(\theta;(x,y)). \tag{9}$$

The intuition is that the perturbation along the gradient norm direction increases the loss value significantly. When training on corrupted distributions, the scarce noisy data scatter sparsely in the high-dimensional space. As a consequence, the neighbor of each datum could not be sufficiently explored, thus producing a sharp loss curve. During test, the unseen noisy data is likely to fall on an unexplored point with a large loss, further causing inaccurate model predictions.

Therefore, instead of directly applying sharpness minimization on the whole dataset, which leads to poor generalization performance Cha et al. (2021) (as demonstrated in Section 4.3), we focus on sharpness minimization over the worst-case distribution Q. By conducting the worst-case sharpness minimization, we can enhance the flatness of our classifier. Consequently, when predicting unknown data during test phase, a flat loss landscape is more likely to produce low loss than a sharp one, hence our SharpDRO can generalize better than other DRO methods. However, the robust performance largely depends on the worst-case distribution Q, so next, we explain our worst-case data selection.

3.2 Worst-Case Data Selection

Generally, the worst-case data selection focuses on finding the most uncertain data distribution Q from the uncertainty set $\mathcal{Q}$, which is a f-divergence ball from the training distribution P (Ben-Tal et al., 2013; Duchi & Namkoong, 2018; Hu et al., 2018). Most works assume each distribution is distinguishable from each other. However, when the distribution index is not available, it would be very hard to select worst-case data. In this section, we investigate two situations: distribution-aware robust generalization and distribution-agnostic robust generalization.

3.2.1 Distribution-Aware Robust Generalization

When given annotations to denote different severity of corruptions, the image data x is paired with class label y and distribution index s. Then, the worst-case distribution Q can be identified as the sub-distribution P_s that yields the largest training error. Hence, we can optimize though:

$$\min_{\theta}\left\{\mathcal{L}_{\text{SharpDRO}} := \max_{P_s \in P}\left\{\mathbb{E}_{(x,y)\sim P_s}\left[\mathcal{L}(\theta;(x,y))\right]\right\} + \mathbb{E}_{(x,y)\sim P_s}\left[\mathcal{R}(\theta;(x,y))\right]\right\}. \tag{10}$$

The first term simply recovers the learning target of GroupDRO (Sagawa et al., 2020; Hu et al., 2018), and the second sharpness minimization term acts as a regularizer. Specifically, we not only emphasize the risk minimization on worst-case distribution P_s, but also enforce low sharpness on P_s. As a result, SharpDRO can learn a flatter loss surface on the worst-case data, thus generalize better compared to GroupDRO, as discussed in Section 4.

3.2.2 Distribution-Agnostic Robust Generalization

Due to the annotations being extremely expensive in real-world applications, a practical challenge is how to learn a robust model without distribution index. Unlike JTT (Liu et al., 2021) which trains the model through two stages, we aim to solve this problem more efficiently by detecting the worst-case data during network training. As the corrupted data essentially lie out-of-distribution from the clean data, so we are motivated to conduct OOD detection (Hendrycks et al., 2019; Liang et al., 2018; Liu et al., 2020; Wei et al., 2022) to find the worst-case data.

Particularly, we re-utilize the previously computed weight perturbation ϵ^* to compute an OOD score:

$$w = \mathbb{E}_{(x,y)\sim P}\left[\max_{i\in\mathcal{Y}} f_i(\theta;(x,y)) - \max_{i\in\mathcal{Y}} f_i(\theta+\epsilon^*;(x,y))\right], \tag{11}$$

where $f(\cdot)$ stands for the c-dimensional label prediction in the label space, whose maximum value is considered as prediction confidence. Intuitively, as the model is much more robust to the clean distribution than the corrupted distribution, the prediction of clean data usually exhibits more stability than scarce noisy data when facing perturbations. Hence, if an example comes from a rarely explored distribution, its prediction certainty would deviate significantly from the original value, thus producing a large OOD score, as shown in Section 4.3. Note that the major difference is that we target generalization on worst-case data, but OOD detection aims to exclude OOD data.

To this end, we can construct our uncertainty set as $\mathcal{Q} := \left\{\sum_{i=1}^n \bar{w}_i \cdot (x_i, y_i) : \bar{w}_i = \frac{w_i}{\frac{1}{n}\sum_{i=1}^n w_i}\right\}$, where normalization on w_i is performed simultaneously. Then, the learning target of our distribution-agnostic SharpDRO is formulated as:

$$\min_\theta \left\{\mathcal{L}_{\text{SharpDRO}} := \sup_{Q\in\mathcal{Q}} \left\{\mathbb{E}_{(x,y)\sim Q}\left[\mathcal{L}(\theta;(x,y))\right]\right\} + \mathbb{E}_{(x,y)\sim Q}\left[\mathcal{R}(\theta;(x,y))\right]\right\}. \tag{12}$$

Therefore, the worst-case data can be selected by focusing on the examples with large OOD scores. In this way, our sharpDRO can be successfully deployed into the distribution-agnostic setting to ensure robust generalization, whose effectiveness is demonstrated by quantitative and qualitative results in Sections 4.2 and 4.3. Next, we give details about implementing SharpDRO.

3.3 Practical Implementation

Overall, the training process of SharpDRO is summarized in Algorithm 1. Note that our SharpDRO requires two backward phases, so the time complexity of this form is twice as much as plain training, for efficient sharpness computation, please refer to (Du et al., 2022; Zhao et al., 2022). In the first step, we record the label prediction p of each data during inference, and simultaneously compute the loss $\mathcal{L}$. Additionally, in the first backward pass, we store the computed gradient $\nabla\mathcal{L}(\theta)$. Further, by adding ϵ^*, we use the perturbed model to compute the second label prediction $\hat{p}$, which is further leveraged to compute the sharpness $\mathcal{R}$. Moreover, in the distribution-agnostic setting, the predictions p and $\hat{p}$ from two forward steps are used to compute the OOD score w. Then, we adding the recorded gradient $\nabla\mathcal{L}(\theta)$ back to the model parameter, and conduct sharpness minimization over the selected worst-case data. In this way, our SharpDRO can be correctly performed. In the next section, we give specific details about our experimental setting and conduct extensive quantitative as well as qualitative analyses to empirically validate the proposed SharpDRO.

Algorithm 1 Training process of SharpDRO

Input: Training set $\mathcal{D}_{\text{train}} = \{x_i, y_i\}_{i=1}^n$ containing Poisson distributed noisy corruptions.
Output: Model parameter θ

for *epoch* $= 1$ to E **do**
  ▷ *First Step*
  Compute loss value $\mathcal{L}(\theta;(x,y))$;
  Compute $\nabla\mathcal{L}(\theta)$ to obtain ϵ^* by first backward pass;
  Compute sharpness via Eq. 9;
  ▷ *Second Step*
  if *Distribution-aware* **then**
    Choose $P_s = \arg\max_{P_s\in P} \mathbb{E}_{(x,y)\sim P_s}\mathcal{L}$ as worst-case distribution;
    Second backward pass by minimizing Eq. 10;
  else if *Distribution-agnostic* **then**
    Compute OOD score w via Eq. 11 and normalize;
    Construct uncertainty set through $Q := \left\{\sum_{i=1}^n \bar{w}_i \cdot (x_i, y_i)\right\}$;
    Second backward pass by minimizing Eq. 12;
  end if
end for

4 Experiments

In our experiments, we first give details about our experimental setup. Then, we conduct quantitative experiments to compare to proposed SharpDRO with the most popular baseline methods by considering both distribution-aware and distribution agnostic settings, which shows the capability of SharpDRO to tackle the most challenging distributions. Finally, we conduct qualitative analyses to investigate the effectiveness of SharpDRO in achieving robust generalization.

Table 1: Quantitative comparisons on distribution-aware robust generalization setting. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Gaussian	ERM	90.9 ± 0.02	89.2 ± 0.02	86.4 ± 0.03	85.9 ± 0.01	83.5 ± 0.01	78.8 ± 0.01
		IRM	91.8 ± 0.01	90.3 ± 0.01	89.5 ± 0.01	86.7 ± 0.02	81.8 ± 0.02	80.0 ± 0.02
		REx	91.3 ± 0.03	89.5 ± 0.02	88.1 ± 0.02	86.7 ± 0.02	83.3 ± 0.01	80.5 ± 0.02
		GroupDRO	90.2 ± 0.03	89.1 ± 0.02	88.4 ± 0.04	84.3 ± 0.01	83.0 ± 0.02	78.2 ± 0.02
		SharpDRO	$\mathbf{92.4 \pm 0.02}$	$\mathbf{91.2 \pm 0.02}$	$\mathbf{90.4 \pm 0.01}$	$\mathbf{88.1 \pm 0.02}$	$\mathbf{86.5 \pm 0.01}$	$\mathbf{82.8 \pm 0.01}$
	Shot	ERM	$\mathbf{92.5 \pm 0.02}$	91.1 ± 0.02	89.9 ± 0.01	85.6 ± 0.03	85.7 ± 0.01	78.8 ± 0.01
		IRM	90.4 ± 0.01	90.3 ± 0.02	89.4 ± 0.02	86.3 ± 0.01	84.3 ± 0.02	79.1 ± 0.02
		REx	91.1 ± 0.02	90.6 ± 0.02	90.2 ± 0.03	86.8 ± 0.02	84.7 ± 0.02	80.5 ± 0.01
		GroupDRO	92.2 ± 0.01	$\mathbf{91.4 \pm 0.01}$	89.4 ± 0.02	84.0 ± 0.01	84.7 ± 0.02	78.3 ± 0.01
		SharpDRO	92.4 ± 0.02	91.1 ± 0.02	$\mathbf{90.3 \pm 0.02}$	$\mathbf{87.5 \pm 0.02}$	$\mathbf{86.4 \pm 0.02}$	$\mathbf{83.3 \pm 0.02}$
	Snow	ERM	90.8 ± 0.01	90.1 ± 0.02	88.1 ± 0.02	88.1 ± 0.02	85.7 ± 0.02	82.6 ± 0.01
		IRM	91.1 ± 0.02	90.7 ± 0.01	89.7 ± 0.02	88.0 ± 0.03	84.6 ± 0.02	83.2 ± 0.03
		REx	91.8 ± 0.02	$\mathbf{91.9 \pm 0.01}$	88.4 ± 0.01	88.3 ± 0.01	88.6 ± 0.01	83.0 ± 0.02
		GroupDRO	91.5 ± 0.02	91.0 ± 0.01	88.7 ± 0.02	88.6 ± 0.02	85.2 ± 0.03	83.5 ± 0.02
		SharpDRO	$\mathbf{93.1 \pm 0.01}$	91.8 ± 0.01	$\mathbf{90.5 \pm 0.02}$	$\mathbf{90.8 \pm 0.02}$	$\mathbf{87.9 \pm 0.01}$	$\mathbf{84.3 \pm 0.02}$
CIFAR100	Gaussian	ERM	68.2 ± 0.01	64.8 ± 0.01	60.6 ± 0.01	56.9 ± 0.01	53.9 ± 0.01	50.2 ± 0.03
		IRM	64.7 ± 0.02	64.7 ± 0.01	62.2 ± 0.01	54.5 ± 0.02	53.4 ± 0.03	50.4 ± 0.01
		REx	68.0 ± 0.03	65.1 ± 0.03	61.8 ± 0.01	56.8 ± 0.01	53.2 ± 0.01	51.5 ± 0.01
		GroupDRO	66.1 ± 0.01	61.7 ± 0.02	59.3 ± 0.03	53.6 ± 0.01	54.0 ± 0.02	50.6 ± 0.02
		SharpDRO	$\mathbf{71.2 \pm 0.02}$	$\mathbf{70.1 \pm 0.01}$	$\mathbf{68.6 \pm 0.01}$	$\mathbf{58.8 \pm 0.01}$	$\mathbf{57.5 \pm 0.02}$	$\mathbf{53.8 \pm 0.03}$
	Shot	ERM	67.6 ± 0.03	65.1 ± 0.01	62.9 ± 0.01	56.0 ± 0.01	55.1 ± 0.01	47.3 ± 0.01
		IRM	67.5 ± 0.02	65.7 ± 0.01	62.7 ± 0.01	59.5 ± 0.01	55.8 ± 0.01	48.3 ± 0.01
		REx	65.7 ± 0.01	63.8 ± 0.02	61.9 ± 0.01	59.3 ± 0.03	53.8 ± 0.01	48.1 ± 0.01
		GroupDRO	67.0 ± 0.02	65.8 ± 0.01	63.1 ± 0.01	58.9 ± 0.01	57.5 ± 0.01	49.3 ± 0.01
		SharpDRO	$\mathbf{69.2 \pm 0.01}$	$\mathbf{67.3 \pm 0.02}$	$\mathbf{65.4 \pm 0.03}$	$\mathbf{62.5 \pm 0.01}$	$\mathbf{57.7 \pm 0.02}$	$\mathbf{50.3 \pm 0.01}$
	Snow	ERM	67.7 ± 0.01	68.1 ± 0.01	64.7 ± 0.01	63.1 ± 0.01	60.5 ± 0.02	57.3 ± 0.01
		IRM	69.3 ± 0.01	67.5 ± 0.02	64.9 ± 0.02	61.0 ± 0.01	58.2 ± 0.01	55.1 ± 0.01
		REx	66.4 ± 0.01	65.9 ± 0.01	62.4 ± 0.01	61.2 ± 0.02	57.5 ± 0.03	56.0 ± 0.02
		GroupDRO	68.0 ± 0.02	68.2 ± 0.01	65.1 ± 0.01	60.9 ± 0.03	59.8 ± 0.01	58.1 ± 0.02
		SharpDRO	$\mathbf{71.5 \pm 0.01}$	$\mathbf{70.8 \pm 0.03}$	$\mathbf{67.5 \pm 0.02}$	$\mathbf{65.5 \pm 0.01}$	$\mathbf{62.3 \pm 0.01}$	$\mathbf{59.2 \pm 0.03}$
ImageNet30	Gaussian	ERM	87.5 ± 0.01	84.6 ± 0.01	81.9 ± 0.01	76.5 ± 0.01	71.2 ± 0.01	65.3 ± 0.01
		IRM	86.6 ± 0.01	84.4 ± 0.03	80.6 ± 0.01	75.2 ± 0.01	70.7 ± 0.03	64.8 ± 0.01
		REx	86.3 ± 0.01	83.8 ± 0.03	81.1 ± 0.02	75.6 ± 0.02	71.5 ± 0.01	66.1 ± 0.03
		GroupDRO	85.1 ± 0.02	84.2 ± 0.01	81.2 ± 0.03	76.3 ± 0.03	72.0 ± 0.02	66.3 ± 0.01
		SharpDRO	$\mathbf{88.4 \pm 0.02}$	$\mathbf{87.6 \pm 0.01}$	$\mathbf{83.3 \pm 0.01}$	$\mathbf{79.1 \pm 0.02}$	$\mathbf{73.6 \pm 0.03}$	$\mathbf{67.7 \pm 0.01}$
	Shot	ERM	86.9 ± 0.01	84.8 ± 0.01	83.6 ± 0.01	79.7 ± 0.01	75.4 ± 0.01	64.6 ± 0.01
		IRM	86.8 ± 0.01	85.1 ± 0.03	81.5 ± 0.01	73.5 ± 0.02	68.5 ± 0.03	62.5 ± 0.03
		REx	83.8 ± 0.01	86.3 ± 0.03	82.5 ± 0.02	73.9 ± 0.01	70.6 ± 0.03	64.0 ± 0.02
		GroupDRO	86.7 ± 0.01	85.6 ± 0.03	84.5 ± 0.01	80.7 ± 0.01	$\mathbf{76.2 \pm 0.04}$	65.4 ± 0.01
		SharpDRO	$\mathbf{88.1 \pm 0.01}$	$\mathbf{87.2 \pm 0.02}$	$\mathbf{84.7 \pm 0.01}$	$\mathbf{82.2 \pm 0.01}$	75.8 ± 0.01	$\mathbf{66.5 \pm 0.02}$
	Snow	ERM	86.7 ± 0.03	85.2 ± 0.01	83.4 ± 0.01	81.1 ± 0.01	75.3 ± 0.01	75.6 ± 0.01
		IRM	85.6 ± 0.01	84.0 ± 0.02	82.1 ± 0.03	79.7 ± 0.01	75.0 ± 0.01	75.6 ± 0.01
		REx	85.4 ± 0.01	84.6 ± 0.02	82.7 ± 0.02	80.5 ± 0.03	75.7 ± 0.03	75.9 ± 0.03
		GroupDRO	86.7 ± 0.01	85.5 ± 0.03	83.4 ± 0.01	81.2 ± 0.02	76.3 ± 0.01	76.6 ± 0.01
		SharpDRO	$\mathbf{88.2 \pm 0.02}$	$\mathbf{88.2 \pm 0.01}$	$\mathbf{85.4 \pm 0.02}$	$\mathbf{81.9 \pm 0.01}$	$\mathbf{79.8 \pm 0.03}$	$\mathbf{79.5 \pm 0.02}$

4.1 Experimental Setup

For distribution-aware situation, we choose GroupDRO (Sagawa et al., 2020), IRM (Arjovsky et al., 2019), REx (Krueger et al., 2021), and ERM for comparisons. As for distribution-agnostic situation, we pick JTT (Liu et al., 2021) and Environment Inference for Invariant Learning (EIIL) (Creager et al., 2021) for baseline methods. For each problem setting, we construct corrupted dataset using CIFAR10/100 (Krizhevsky et al., 2009) and ImageNet30 (Russakovsky et al., 2015) datasets. Specifically, we following (Hendrycks & Dietterich, 2019) to perturb the image data with severity level varies from 1 to 5 by using three types of corruptions: "Gaussian Noise", "Shot Noise", and "Snow". Moreover, the clean data are considered as having a corruption severity of 0. For each corrupted distribution, we sample them with different probabilities by following Poisson distribution $P(s; \lambda = 1)$, *i.e.*, for s varies from 0 to 5, the sample probabilities are $\{0.367, 0.367, 0.184, 0.061, 0.015, 0.003\}$, respectively. Then, we test the robust performance on each data distribution. For hyper-parameter ρ, we follow (Foret et al., 2020) by setting it to 0.05 to control the magnitude of ϵ^*. For each experiment, we conduct three independent trials and report the average test accuracy with standard deviations.

4.2 Quantitative Comparisons

In our quantitative comparisons, we focus on three questions: 1) Can SharpDRO perform well on two situations of robust generalization? 2) Does SharpDRO generalize well on the most severely corrupted distributions? and 3) Is SharpDRO able to tackle different types of corruption? To answer these questions, we conduct experiments on both two settings by testing on different levels of severity. Moreover, we consider three types of corruption.

Table 2: Quantitative comparisons on distribution-agnostic robust generalization setting. Averaged accuracy (%) with standard deviations are computed over three independent trails.

Dataset	Type	Method	Corruption Severity 0	1	2	3	4	5
CIFAR10	Gaussian	JTT	89.9 ± 0.02	88.8 ± 0.02	$86.5 \pm 0..02$	86.1 ± 0.02	83.4 ± 0.03	79.8 ± 0.02
		EIIL	88.6 ± 0.02	87.5 ± 0.03	86.3 ± 0.03	85.4 ± 0.02	83.2 ± 0.03	78.8 ± 0.01
		SharpDRO	$\mathbf{91.3 \pm 0.01}$	$\mathbf{90.2 \pm 0.02}$	$\mathbf{88.7 \pm 0.01}$	$\mathbf{86.1 \pm 0.02}$	$\mathbf{84.2 \pm 0.02}$	$\mathbf{82.7 \pm 0.02}$
	Shot	JTT	91.3 ± 0.02	90.5 ± 0.03	89.3 ± 0.01	86.5 ± 0.02	83.1 ± 0.02	79.8 ± 0.02
		EIIL	90.3 ± 0.03	90.1 ± 0.02	88.3 ± 0.01	86.2 ± 0.02	82.3 ± 0.03	78.5 ± 0.02
		SharpDRO	$\mathbf{91.6 \pm 0.01}$	$\mathbf{90.5 \pm 0.02}$	$\mathbf{89.8 \pm 0.02}$	$\mathbf{87.1 \pm 0.01}$	$\mathbf{85.3 \pm 0.02}$	$\mathbf{81.7 \pm 0.01}$
	Snow	JTT	88.6 ± 0.02	87.8 ± 0.03	86.5 ± 0.02	87.2 ± 0.02	84.2 ± 0.02	83.2 ± 0.03
		EIIL	88.3 ± 0.02	87.8 ± 0.01	85.6 ± 0.02	87.3 ± 0.03	85.2 ± 0.04	82.3 ± 0.01
		SharpDRO	$\mathbf{91.6 \pm 0.01}$	$\mathbf{91.1 \pm 0.02}$	$\mathbf{90.8 \pm 0.01}$	$\mathbf{89.7 \pm 0.02}$	$\mathbf{86.2 \pm 0.01}$	$\mathbf{83.8 \pm 0.02}$
CIFAR100	Gaussian	JTT	68.0 ± 0.02	65.3 ± 0.02	61.3 ± 0.01	56.3 ± 0.01	54.2 ± 0.03	51.2 ± 0.02
		EIIL	67.2 ± 0.01	66.2 ± 0.02	61.0 ± 0.02	55.8 ± 0.02	54.6 ± 0.03	52.1 ± 0.02
		SharpDRO	$\mathbf{69.6 \pm 0.03}$	$\mathbf{68.0 \pm 0.02}$	$\mathbf{63.6 \pm 0.03}$	$\mathbf{58.2 \pm 0.02}$	$\mathbf{55.6 \pm 0.03}$	$\mathbf{52.4 \pm 0.03}$
	Shot	JTT	66.3 ± 0.02	65.3 ± 0.03	63.4 ± 0.02	56.6 ± 0.04	55.5 ± 0.04	48.6 ± 0.04
		EIIL	66.5 ± 0.02	65.3 ± 0.03	62.8 ± 0.04	57.5 ± 0.02	56.5 ± 0.01	49.5 ± 0.01
		SharpDRO	$\mathbf{68.9 \pm 0.02}$	$\mathbf{66.2 \pm 0.03}$	$\mathbf{64.9 \pm 0.03}$	$\mathbf{59.8 \pm 0.02}$	$\mathbf{56.5 \pm 0.03}$	$\mathbf{51.0 \pm 0.02}$
	Snow	JTT	67.5 ± 0.01	68.1 ± 0.02	65.3 ± 0.02	64.3 ± 0.02	60.2 ± 0.02	57.8 ± 0.02
		EIIL	68.2 ± 0.03	69.1 ± 0.03	65.2 ± 0.02	64.0 ± 0.02	61.0 ± 0.04	57.5 ± 0.04
		SharpDRO	$\mathbf{70.6 \pm 0.02}$	$\mathbf{69.9 \pm 0.03}$	$\mathbf{66.7 \pm 0.03}$	$\mathbf{64.4 \pm 0.02}$	$\mathbf{61.9 \pm 0.03}$	$\mathbf{60.7 \pm 0.03}$
ImageNet30	Gaussian	JTT	87.3 ± 0.02	84.5 ± 0.02	82.3 ± 0.04	75.6 ± 0.01	72.1 ± 0.04	66.5 ± 0.02
		EIIL	$\mathbf{88.2 \pm 0.02}$	85.2 ± 0.03	81.3 ± 0.02	74.5 ± 0.02	71.5 ± 0.02	65.0 ± 0.04
		SharpDRO	87.1 ± 0.02	$\mathbf{86.9 \pm 0.02}$	$\mathbf{83.5 \pm 0.03}$	$\mathbf{78.0 \pm 0.02}$	$\mathbf{72.2 \pm 0.02}$	$\mathbf{66.6 \pm 0.03}$
	Shot	JTT	86.5 ± 0.02	85.4 ± 0.03	82.6 ± 0.04	79.6 ± 0.04	$\mathbf{77.2 \pm 0.04}$	65.0 ± 0.01
		EIIL	85.5 ± 0.01	86.3 ± 0.04	81.6 ± 0.02	80.2 ± 0.03	75.3 ± 0.02	64.4 ± 0.03
		SharpDRO	$\mathbf{86.8 \pm 0.02}$	$\mathbf{87.2 \pm 0.03}$	$\mathbf{83.2 \pm 0.03}$	$\mathbf{81.4 \pm 0.06}$	76.6 ± 0.03	$\mathbf{65.3 \pm 0.03}$
	Snow	JTT	86.0 ± 0.04	85.8 ± 0.02	82.3 ± 0.03	80.4 ± 0.02	74.6 ± 0.02	73.5 ± 0.02
		EIIL	87.5 ± 0.01	85.4 ± 0.02	83.5 ± 0.04	$\mathbf{81.6 \pm 0.01}$	76.3 ± 0.01	75.8 ± 0.02
		SharpDRO	$\mathbf{87.5 \pm 0.03}$	$\mathbf{86.7 \pm 0.02}$	$\mathbf{85.4 \pm 0.02}$	81.5 ± 0.03	$\mathbf{78.9 \pm 0.02}$	$\mathbf{78.5 \pm 0.03}$

Distribution-Aware Robust Generalization As shown in Table 1, we can see that SharpDRO surpasses other methods with larger performance gains as the corruption severity goes stronger. Especially, in ImageNet30 dataset on "snow" corruption, improvement margin between SharpDRO and second-best method is 1.5% with severity of 0, which is further increased to about 3% with severity of 0, which indicates the capability of SharpDRO on generalization against severe corruptions. Moreover, SharpDRO frequently outperforms other methods on all three types of corruption, which manifests the general robustness of SharpDRO against various corruption types.

Distribution-Agnostic Robust Generalization As shown in Table 2, we can see a similar phenomenon in Table 1 that the more severe corruptions are applied, the larger performance gains SharpDRO achieves. Especially, in the CIFAR10 dataset corrupted by "Gaussian Noise", SharpDRO shows about 1.4% performance gains upon the second-best method with severity 0, which is further increased to almost 3% with severity 5. Moreover, SharpDRO is general to all three corruption types, as it surpasses other methods in most cases. Therefore, the proposed method can perfectly generalize to worst-case data even without the distribution annotations.

4.3 Qualitative Analysis

To investigate the effectiveness of SharpDRO, we first conduct an ablation study to show that the Sharpness minimization on the worst-case data is essential for achieving generalization with robustness. Then, we utilize the gradient norm, an important criterion to present training stability, to validate that our method is stable for severely corrupted distributions. Then, we analyze the hyper-parameter ρ and OOD score $\bar{w}$ to disclose the effectiveness of sharpness minimization and worst-case data selection. All analyses are conducted using CIFAR10 with "Gaussian Noise" corruption.

Ablation Study By eliminating the worst-case data selection, we recover the original sharpness minimization method, which is denoted as SAM Foret et al. (2020). Then, we remove the sharpness minimization module, which is basically training via GroupDRO. The ablation results are shown in Table 3. We can see that deploying SAM on the whole training dataset can achieve improved results on the clean dataset. However, the robust performance on corrupted distributions are even worse than GroupDRO.

Table 3: Ablation study.

Method	Corruption Severity 0	1	2	3	4	5
w/o worst-case data selection	93.2	90.5	87.6	82.1	80.5	75.4
w/o sharpness minimization	90.2	89.1	88.4	84.3	83.0	78.2
SharpDRO	92.4	91.2	90.4	88.1	86.5	82.8

This could be because that sharpness is easy to be dominated by principle distributions, which is misleading for generalization to small distributions. Thus, the sharpness of corrupted data would be sub-optimal. As for GroupDRO, it fails to produce a flat loss surface for worst-case data, hence cannot generalize as well as the proposed SharpDRO.

Distributional Stability To show our method can be stable even in the most challenging distributions, we show the gradient norm on a validation set including corruption severity from 0 to 5. As shown in Figure 4, SharpDRO not only produces the smallest norm value but also can ensure almost equal gradient norm across all corrupted distributions, which indicates that SharpDRO is the most distributionally stable method among all compared methods.

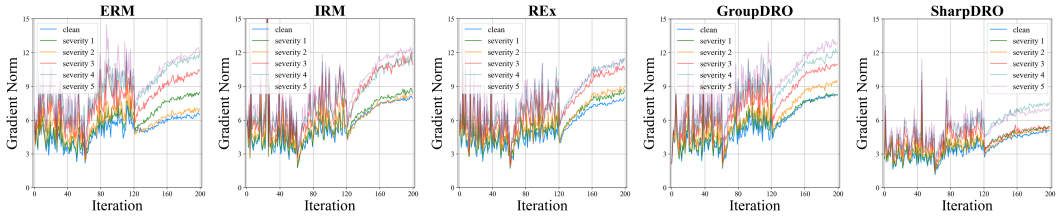

Figure 4: Gradient norm comparisons between different methods over all corrupted distributions.

Parameter Analysis To understand how the scale parameter ρ affects our generalization performance, we conduct sensitivity analysis by changing this value and show the test results of different distributions. In figure 5, we find an interesting discovery that as ρ increases, which indicates the perturbation magnitude ϵ^* enlarges, would enhance the generalization of severely corrupted data but degrades the performance of slightly corrupted data. This might be because the exploration of hard distributions needs to cover wide range of neighborhood to ensure generalization. On the contrary, exploration too far on easy distributions can reach out-of-distribution, thus causing performance degradation. Therefore, for practitioners who aim to generalize on small and difficult datasets, we might be able to enhance performance by aggressively setting a large perturbation scale.

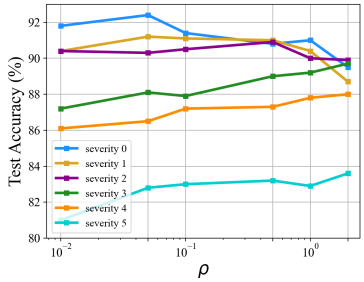

Figure 5: Parameter sensitivity of ρ whose value is set to $\{0.01, 0.05, 0.1, 0.5, 1, 2\}$.

OOD Score Analysis The OOD score is leveraged to select worst-case data for the distribution-agnostic setting. To show its effectiveness in selecting the noisy data, we plot the value distribution of OOD scores from all corrupted distributions in Figure 6. We can see the tendency that more severely corrupted data to have larger OOD scores. Therefore, our OOD score is a valid criterion to select worst-case data. Note that during the training process, the worst-case data would be gradually learned, thus the OOD score can become smaller, which explains why the value distribution of our score is not as separable as OOD detection does.

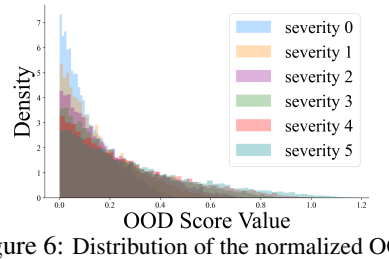

Figure 6: Distribution of the normalized OOD score $\bar{w}$ on clean distribution ($s = 0$), and corrupted distribution from $s = 1$ to 5. Values are selected in epoch 30.

5 Conclusion

In this paper, we proposed a SharpDRO approach to enhance the generalization performance of DRO methods. Specifically, we focus on minimizing the sharpness of worst-case data to learn flat loss surfaces. As a result, SharpDRO is more robust to severe corruptions compared to other methods. Moreover, we apply SharpDRO to distribution-aware and distribution-agnostic settings and proposed an OOD detection process to select the worst-case data when the distribution index is not known. Extensive quantitative and qualitative experiments have been conducted to show that SharpDRO can deal with the most challenging corrupted distributions and achieve improved generalization results compared to well-known baseline methods.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22405–22418, 2021.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *ICLR*, 2017.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, pp. 2189–2200. PMLR, 2021.

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent YF Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, pp. 1929–1938. PMLR, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *ICML*, pp. 2029–2037. PMLR, 2018.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pp. 5815–5826. PMLR, 2021.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, pp. 6781–6792. PMLR, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, volume 33, pp. 21464–21475, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pp. 97–105. PMLR, 2015.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, volume 29, 2016.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Focus on the common good: Group distributional robustness follows. In *ICLR*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pp. 7167–7176, 2017.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, volume 33, pp. 2958–2969, 2020.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *ICML*, pp. 12345–12355. PMLR, 2021.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *ICML*, 2022.

Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *CVPR*, pp. 8156–8165, 2021.