Deep Diffusion-Invariant Wasserstein Distributional Classification

We would like to thank the reviewers for their time and effort to read our paper and provide constructive suggestions. We carefully addressed all comments as closely as possible. All reviewers agree that the overall system contains a number of potentially interesting ingredients and can represent a worthy contribution. Hopefully, the paper in its present form satisfies the requirements of the reviewers. Thus, please give a chance to publish our work in NeurIPS 2020.

Reviewer $\#1$

Some details were omitted in the main paper due to the lack of space, but we attempted to add as many details as possible to Appendix. According to your instruction, we will move the core definitions in Appendix to the main paper. Please appreciate in-depth and theoretical analysis of the proposed method in the main paper and Appendix.

C1. Gaussian measures. We represented target Gaussian measures as r-rank covariance matrices, $\Sigma_c = \mathbf{M}_c^T \mathbf{M}_c$, where $\mathbf{M}_c$ denotes $(r \times d)$ size of random matrix for the c-th class and all indices are i.i.d uniform random variables. We use the centered Gaussian measures, because it is stationary against the diffusion operator in Proposition 1. If we consider non-null means, the Mehler's formula is not explicitly defined, which is a crucial problem for efficient learning.

C2. Details. The batch-size was set to 128. We will more rigorously define the notations in the camera ready paper.

Reviewer $\#2$

C1. Our motivation on the data with multiple locations. We mean that each pixel of data can change its location, if a different perturbation is applied. For example, a single 2D image is represented as a set of the pixels, $\{x \in \mathbb{R}^3\}$, where x denotes a single pixel. Then, each pixel x can be contaminated into by various factors (*e.g.*, local noises and rotations), which produces multiple variants of x (*e.g.*, $x + \varepsilon_1$ and $x + \varepsilon_2$), as depicted by red points in Figure 1. If $\varepsilon_1 = (-10, 0, 0)$, *i.e.*, random noise, x is shifted to the left along the x-axis, which has a location $x + (-10, 0, 0)$. The CIFAR-10-C dataset contains the aforementioned variations of a single data. *Our goal is to cover all of these variants by finding the optimal representation in terms of the Wasserstein uncertainty and to classify them into the same class.*

C2. Our motivation on Wasserstein uncertainty. Our motivation is to suggest a new framework to deal with *stochastic perturbations* with tools developed in OT. The Wasserstein uncertainty corresponds to the 2-Wasserstein distance between diffusion variant (affected by perturbations) and invariant (hardly affected by perturbations) label measures. Thus, minimizing the Wasserstein uncertainty is equivalent to protecting the proposed label estimation process from several perturbations. The Wasserstein distance was used, because it can handle measures (*i.e.*, distributions).

C3. Real-world datasets. We have results on real-world perturbations and will include them to the final version.

C4. Table 1. We believe that a severe rotation setting, *i.e.*, $\theta_2 = 2\pi$, makes our diffusion classifier g much easier to capture global information, in which the measure ν becomes smoother and predictable in the Wasserstein space.

Reviewer $\#3$

C1. Why diffusion-invariance help to improve robustness? The goal of conventional classification methods is to fit the mapped data $f(x) \in \mathbb{R}^d$ to the true label $\hat{y} \in \mathbb{R}^d$ given as an one-hot vector. Contrary, the goal of our method is to fit the mapped distribution $\left\{f(x) \in \mathbb{R}^d\right\}$ to a set of i.i.d Gaussian vectors $\left\{Z \in \mathbb{R}^d\right\}$ given as a probability measure, where the Gaussian probability measure is robust to random perturbations (*i.e.*, stationary against the diffusion operator in Proposition 1). Thus, the shape of $\left\{f(x) \in \mathbb{R}^d\right\}$ should be similar to that of the target Gaussian distribution. In this context, the diffusion-invariance term in eq.(4) measures how $\left\{f(x) \in \mathbb{R}^d\right\}$ is differed from the target Gaussian distribution. If the geometric transformations are severe, the shape of $\left\{f(x) \in \mathbb{R}^d\right\}$ is largely fluctuated and is highly non-Gaussian, which makes the diffusion-invariance term to yield a large value. Then, our method tries to minimize this diffusion-invariance term, and thus can reduce the aforementioned fluctuation (*i.e.*, perturbation).

C2. How much distortion can be tolerated? Corollary 1 shows that the probability of accurate classification is bounded exponentially inverse to the value of δ, which is the diffusion-invariance term (degree of distortion) in eq.(4).

C3. Balance between invariance and sensitivity Our method can balance between the invariance to image distortion (by diffusion invariance term) and sensitivity to image content change (by intrinsic distance term). We will Clarify it.

C4. Why g^ϑ is maximized? The choice of function g in eq.(4) is crucial for the diffusion-invariance term. By maximizing g^ϑ, we can find the most sensitive function which will accurately measure the diffusion-invariance.

Reviewer $\#4$

We tried to describe the global picture in Figure 1 and provide details in Appendix. We will further detail our method.

C1. Explanation on performance gap on CIFAR10-C. The sampled images in the CIFAR-10-C only possess 19 pre-fixed deterministic variations, which induce low randomness of perturbed objects. While the main contribution of our method is to show not only the robustness to severe perturbations but also the robustness to highly stochastic perturbations, theoretical advantages developed in Section 3.3 are limited due to the low randomness, which leads a small margin in accuracy. As we mentioned above, the perturbed images are recognized as: $X^\varepsilon = \{(x + \varepsilon) \in \mathbb{R}^3\}$. In this case, the probability measure for random variable ε can be defined as $\sum_{n=1}^{19} \delta_n$ for Dirac-delta measure δ supporting 19 varying images. However, this probability measure is highly concentrated and non-smooth, which can not be properly captured by our diffusion classifier g.