Interpolation for Robust Learning: Data Augmentation on Wasserstein Geodesics

Jiacheng Zhu [1] **Jielin Qiu** [1] **Aritra Guha** [2] **Zhuolin Yang** [3] **XuanLong Nguyen** [4] **Bo Li** [3] **Ding Zhao** [1]

Abstract

We propose to study and promote the robustness of a model as per its performance through the interpolation of training data distributions. Specifically, (1) we augment the data by finding the worst-case Wasserstein barycenter on the geodesic connecting subpopulation distributions of different categories. (2) We regularize the model for smoother performance on the continuous geodesic path connecting subpopulation distributions. (3) Additionally, we provide a theoretical guarantee of robustness improvement and investigate how the geodesic location and the sample size contribute, respectively. Experimental validations of the proposed strategy on *four* datasets, including CIFAR-100 and ImageNet, establish the efficacy of our method, e.g., our method improves the baselines' certifiable robustness on CIFAR10 up to 7.7%, with 16.8% on empirical robustness on CIFAR-100. Our work provides a new perspective of model robustness through the lens of Wasserstein geodesic-based interpolation with a practical off-the-shelf strategy that can be combined with existing robust training methods.

1. Introduction

Deep neural networks (DNNs) have shown tremendous success in an increasing range of domains such as natural language processing, image classification & generation and even scientific discovery (e.g. (Bahdanau et al., 2014; Krizhevsky et al., 2017; Ramesh et al., 2022)). However, despite their super-human accuracy on training datasets, neural networks may not be robust. For example, adding imperceptible perturbations, e.g., *adversarial attacks*, to the inputs may cause neural networks to make catastrophic mistakes (Szegedy et al., 2014; Goodfellow et al., 2014). Conventional defense techniques focus on obtaining adversarial perturbations (Carlini & Wagner, 2017; Goodfellow et al., 2014) and augmenting them to the training process. For instance, the projected gradient descent (PGD) (Madry et al., 2018), which seeks the worst-case perturbation via iterative updates, marks a category of effective defense methods.

Recent works show that additional training data, including data augmentation and unlabeled datasets, can effectively improve the robustness of deep learning models (Volpi et al., 2018; Rebuffi et al., 2021). Despite augmenting worst-case samples with gradient information, an unlabeled, even out-of-distribution dataset may also be helpful in this regard (Carmon et al., 2019; Bhagoji et al., 2019). Among available strategies, Mixup (Zhang et al., 2018), which interpolates training samples via convex combinations, is shown to improve both the robustness and generalization (Zhang et al., 2021). Moreover, Gaussian noise is appended to training samples to achieve certifiable robustness (Cohen et al., 2019a), which guarantees the model performance under a certain degree of perturbation. However, most aforementioned approaches operate on individual data samples (or pairs) and require the specification of a family of augmentation policies, model architecture, and additional datasets. Thus, the robustness can hardly be generalizable, e.g., to out-of-sample data examples. To this end, we are curious: *Can we let the underlying data distribution guide us to robustness?*

This paper proposes a framework to augment synthetic data samples by interpolating the underlying distributions of the training datasets. Specifically, we find the *worst-case interpolation distributions* that lie on the decision boundary and improve the model's smoothness with samples from these distributions. This is justified by the fact that better decision margin (Kim et al., 2021a) and boundary thickness (Moosavi-Dezfooli et al., 2019) are shown to benefit robustness.

The probabilistic perspective allows us to formulate the distribution interpolation as barycenter distribution on *Wasserstein geodesics*, where the latter can be reviewed as a dynamic formulation of optimal transport (OT) (Villani, 2009). This formulation has the following benefits: (a). It nicely connects our worst-case barycenter distribution to distributionally robust optimization (Duchi & Namkoong, 2021)

[1]Carnegie Mellon University [2]Data Science & AI Research, AT&T Chief Data Office [3]University of Illinois at Urbana-Champaign [4]University of Michigan Ann Arbor. Correspondence to: Jiacheng Zhu <jzhu4@andrew.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

on the geodesic and substantiates our method's benefit in promoting robustness. (b). The geodesic provides a new protocol to assess one model's robustness beyond the local area surrounding individual data points. (c). Our data augmentation strategy also exploits the local distribution structure by generalizing the interpolation from (mostly Euclidean) feature space to probability space. (d). Apart from most OT use cases that rely on coupling, we reformulate the objective to include explicit optimal transport maps following McCann's interpolation (McCann, 1997). This enables us to deal with out-of-sample data and appreciate the recent advances in OT map estimation (Perrot et al., 2016; Zhu et al., 2021; Amos, 2022), neural OT map estimators (Makkuva et al., 2020; Seguy et al., 2017; Fan et al., 2021) for large-scale high-dimensional datasets, and OT between labeled datasets (Alvarez-Melis & Fusi, 2020; Fan & Alvarez-Melis, 2022).

1.1. Related Work

Interpolation of data distributions (Gao & Chaudhari, 2021) propose a transfer distance based on the interpolation of tasks to guide transfer learning. Meta-learning with task interpolation (Yao et al., 2021), which mixes features and labels according to tasks, also effectively improves generalization. For gradual domain adaptation, (Wang et al., 2022) interpolate the source dataset towards the target in a probabilistic fashion following the OT schema. Moreover, the interpolation of environment distribution (Huang et al., 2022) is also effective in reinforcement learning. The recently proposed rectified flow (Liu et al., 2022) facilities a generating process by interpolating data distributions in their optimal paths. In addition, the distribution interpolant (Albergo & Vanden-Eijnden, 2022) has motivated an efficient flow model which avoids the costly ODE solver but minimizes a quadratic objective that controls the Wasserstein distance between the source and the target. Recently, the interpolation among multiple datasets realized on generalized geodesics (Fan & Alvarez-Melis, 2022) is also shown to enhance generalizability for pretraining tasks. Most aforementioned works focus on the benefit of distribution interpolation in generalization.

On the other hand, Mixup (Zhang et al., 2018), which augments data by linear interpolating between two samples, is a simple yet effective data augmentation method. As detailed in survey (Cao et al., 2022; Lewy & Mańdziuk, 2022), there are plenty of extensions such as CutMix (Yun et al., 2019), saliency guided (Uddin et al., 2020), AugMix (Hendrycks et al., 2019), manifold mixup (Verma et al., 2019), and so on (Yao et al., 2022). A few studies have explored the usage of optimal transport (OT) ideas within mixup when interpolating features (Kim et al., 2020; 2021b). However, those methods focus on individual pairs, thus neglecting the local distribution structure of the data. One recent work (Greenewald et al., 2021) also explores mixing multiple-batch samples with OT alignment. Although their proposed K-mixup better preserves the data manifold, our approach aims to determine the worst-case Wasserstein barycenter, achieved through an interpolation realized by transport maps.

Data augmentation for robustness. Augmentation of data(Rebuffi et al., 2021; Volpi et al., 2018) or more training data (Carmon et al., 2019; Sehwag et al., 2021) can improve the performance and robustness of deep learning models. However, (Schmidt et al., 2018) show that sample complexity for robust learning may be prohibitively large when compared to standard learning. Moreover, with similar theoretical frameworks, recent papers (Deng et al., 2021; Xing et al., 2022) further establish theoretical justifications to characterize the benefit of additional samples for model robustness. However, additional data may not be available; in this work, we therefore use a data-dependent approach to generate additional data samples.

1.2. Contributions

Our key contributions are as follows. We propose a data augmentation strategy that improves the robustness of the label prediction task by finding worst-case data distributions on the interpolation within training distributions. This is realized through the notion of Wasserstein geodesics and optimal transport map, and further strengthened by connection to DRO and regularization effect. Additionally, we also provide a theoretical guarantee of robustness improvement and investigate how the geodesic location and the sample size contribute, respectively. Experimental validations of the proposed strategy on *four* datasets including CIFAR-100 and ImageNet establish the efficacy of our method.

Organization. The remainder of this paper is organized as follows. Section 2 provides the necessary background about distributional robustness and optimal transport distances. In Section 3, we introduce the interpolation strategy for robustifying a classification task. Connections to DRO(Distributionally Robust Optimization) setting further help ground our strategy. Additionally, we also extend this to a regularization framework that generalizes the data interpolation approach. Theoretical analysis for the data augmentation strategy is provided in Section 4, while Section 5 provides experimental validation justifying our approach.

2. Preliminaries

Consider a classification problem on the data set $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y} \subseteq \mathbb{R}^y$ are drawn i.i.d from a joint distribution $\mathcal{P}^{all}_{x,y}$. Having a loss criteria $l(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, the task is to seek a prediction function $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}^y$ that minimizes the standard loss $L(\theta) = \mathbb{E}_{x,y\sim\mathcal{P}^{all}_{x,y}}[l(f_\theta(x), y)]$, and in practice people optimize the empirical loss $\hat{L}(\theta) = 1/n \sum_{i=1}^n l(f(x_i), y_i)$ following the Empirical Risk Minimization (ERM). In large-

scale classification tasks such as k-class image classification, the label is the one-hot encoding of the class as $y_i \in \{0,1\}^k := \mathbb{R}^k$ and l are typically cross-entropy loss.

Adversarial training & distributional robustness Typically, adversarial training is a minimax optimization problem (Madry et al., 2018) which finds adversarial examples $x+\delta$ within a perturbation set $\mathcal{S} = \{\delta : \|\delta\|_\infty \leq \epsilon, \epsilon > 0\}$. While finding specific attack examples is effective, there are potential issues such as overfitting on the attack pattern (Kurakin et al., 2016; Xiao et al., 2019; Zhang & Wang, 2019).

An alternative approach is to capture the distribution of adversarial perturbations for more generalizable adversarial training (Dong et al., 2020). In particular, the optimization problem solves a distributional robust optimization (Duchi & Namkoong, 2021; Weber et al., 2022) as follows:

$$\min_{\theta} \sup_{Q_{x,y} \in \mathcal{U}_P} \mathbb{E}_{x,y\sim Q_{x,y}}[l(f_\theta(x), y)], \tag{1}$$

where $\mathcal{U}_P \subseteq \mathcal{P}(\mathcal{Z})$ is a set of distributions with constrained support. Intuitively, it finds the worst-case optimal predictor f_θ^* when the data distribution P is perturbed towards an adversarial distribution $\mathcal{U}_P$. However, the adversarial distribution family needs to be properly specified. In addition, adversarial training may still have unsmooth decision boundaries (Moosavi-Dezfooli et al., 2019), thus, are vulnerable around unobserved samples.

Optimal transport Given two probability distributions $\mu, \nu \in \mathcal{M}(\mathcal{X})$, where $\mathcal{M}(\mathcal{X})$ is the set of Borel measures on $\mathcal{X}$. The optimal transport (OT) (Villani, 2009) finds the optimal joint distribution and quantifies the movement between μ and ν. Particularly, the Wasserstein distance is formulated as $W_p(\mu,\nu) := (\inf_{\pi\in\Pi} \int_{\mathcal{X}\times\mathcal{X}} d^p(x,y) d\pi(x,y))^{1/p}$, where $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ is the ground metric cost function, and Π denotes set of all joint probability measures on $\mathcal{X} \times \mathcal{X}$ that have the corresponding marginals μ and ν.

Wasserstein barycenter and geodesic Equipped with the Wasserstein distance, we can average and interpolate distributions beyond the Euclidean space. The Wasserstein barycenter $\mu_\alpha^{\{\nu_i\}}$ of a set of measures $\{\nu_1, ..., \nu_N\}$ in a probability space $\mathbb{P} \subset \mathcal{M}(\mathcal{X})$ is a minimizer of objective U_{wb}^N over $\mathbb{P}$, where

$$U_{wb}^N(\mu) := \sum_{i=1}^N \alpha_i W(\mu, \nu_i), \tag{2}$$

and α_i are the weights such that $\sum \alpha_i = 1$ and $\alpha_i > 0$. Let ν_0 and ν_1 be two measure with an existing optimal transportation map, T, satisfying $\nu_1 = T\#\nu_0$ and $W_2^2(\nu_0,\nu_1) = \int_{\mathcal{X}} \|x - T(x)\|^2 d\nu_0$. Then, for each t, the barycenter distribution(denoted by μ_t) corresponding to $\alpha_0 = t, \alpha_1 = 1-t$ in Eq.(2), lies on the geodesic curve connecting ν_0 and ν_1 . Moreover, (McCann, 1997),

$$\mu_t := ((1-t)\mathrm{Id} + tT)\#\nu_0, \tag{3}$$

where Id is the identity map. While the above McCann's interpolation is defined only for two distributions, the *Wasserstein barycenter* eq. (2) can be viewed as a generalization to $N \geq 2$ marginals. To this point, we are able to look for adversarial distributions leveraging such interpolation.

3. Adversarial robustness by interpolation

For a k-classification task, it is natural to view the data of each category are samples from distinct distributions $P_{x,y}^{all} = \sum_i^k P_{x,y}^i$. Following eq.(1), we can write the following minimax objective

$$\min_f \max_\alpha \mathbb{E}_{x,y\sim U^{wb}(\alpha)}\left[l(f(x), y)\right] \tag{4}$$

where $U^{wb}(\alpha)$ is the Wasserstein barycenter (eq. 2), in other words, the interpolation of different category distributions. In fact, this corresponds to a special case of DRO eq.(1) as shown in the below proposition, and naturally corresponds to the adversarial distribution within a particular Wasserstein ball, for distributional robustness.

Proposition 3.1. *Suppose the original data $\{X_i, Y_i,\}_{i=1}^{n_0}$ are iid with distribution satisfying $X|Y = y \sim \nu_\mu(X|Y) = N_d(y\mu, \sigma^2 I)$, $Y \sim \nu(Y) = U\{-1,1\}$. Assume the loss function is given by $l(f(x), y) = \mathbb{1}(f(x) \neq y)$. We consider $f(x) = f_{\alpha,\beta}(x)$ of the form sign$(\alpha' x + \beta)$.Then the following holds*

$$\begin{aligned} & \max_{A(\nu_\mu,\epsilon)} \inf_{\alpha,\beta} \mathbb{E}_{(X,Y)\sim\eta(X|Y)\nu(Y)} l(f_{(\alpha,\beta)}(X), Y) \\ = \; & inf_{\alpha,\beta} \max_{A(\nu_\mu,\epsilon)} \mathbb{E}_{(X,Y)\sim\eta(X|Y)\nu(Y)} l(f_{(\alpha,\beta)}(X), Y) \\ = \; & \mathbb{E}_{(X,Y)\sim\nu_{(1-\epsilon)\mu}(X|Y)\nu(Y)} l(f_{(\mu,0)}(X), Y). \end{aligned} \tag{5}$$

Here, $A(\nu_\mu, \epsilon) = \{\eta(X|Y) : \eta(\cdot|Y)$ is Gaussian, $W(\eta(X|Y), \nu_\mu(X|Y)) \leq \epsilon, \forall Y \in \{-1,1\}\}$.

This proposition, whose proof is discussed in Section A.4.

3.1. Data augmentation: worst-case interpolation

In this section, we proceed to device computation strategies for the worst-case barycenter approach. Akin to robust training methods that seek adversarial examples approximately, such as PGD (Madry et al., 2018) and FGSM (Goodfellow et al., 2014), our paradigm learns an adversarial distribution (Dong et al., 2020) and then samples from these distributions.

Here we make an important choice by focusing on the interpolation between $K = 2$ distributions rather than a generalized K-margin barycenter. To date, efficient computational solutions for for multimarginal OT (MOT) are still nascent (Lin et al., 2022). Solving an MOT with linear

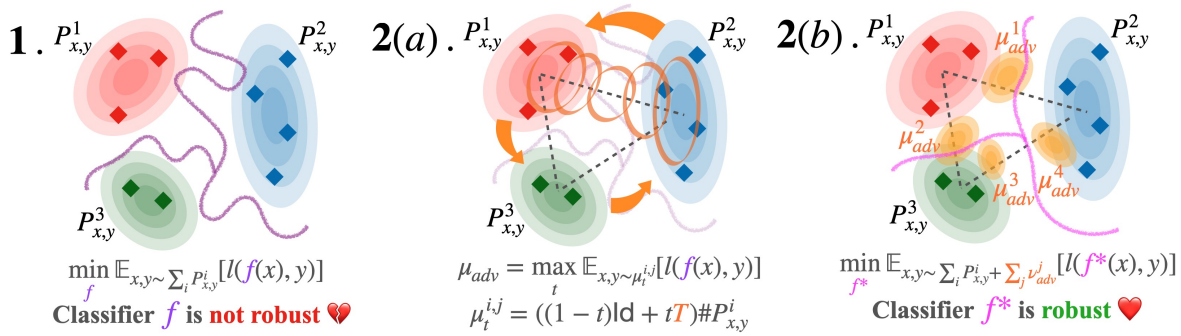


Figure 1: **1**: For a classification problem, the data samples give empirical access to the underlying data distribution, which is a mixture of distinct subpopulation distributions. **2(a)**: The inner maximization step finds the worst-case Wasserstein barycenter on the geodesic. **2(b)**: The outer minimization step updates the predictive function with augmented samples.

programming would scale to $O(n^K)$, and would be more costly when we want to solve for (free-support) barycenters to facilitate robustness in deep learning pipelines. Focusing on $K = 2$ not only (1) allows us to focus on the decision boundary between the different categories of data, and even to deepen our understanding with theoretical analysis for robustness. (2) Achieve computation efficiency by leveraging established OT algorithms when iterating between pairs of distributions, as illustrated in Fig.(1).

Thus, our data augmentation strategy is to sample data from the worst-case adversarial Wasserstein barycenter $\{\hat{x}_i, \hat{y}_i\} \sim \mu_{adv}^{\{P_0,P_1\}}$, which can be obtained as an inner maximization problem,

$$\mu_{adv}^{\{P_0,P_1\}} := \max_{t\in[0,1]} \mathbb{E}_{x,y\sim\mu_t^{\{P_0,P_1\}}}\left[l(f_\theta(x),y)\right]. \tag{6}$$

Despite a list of recent advances (Villani, 2009; Korotin et al., 2022a; Fan et al., 2020), the usage of Wasserstein barycenter may still be limited by scalability, and those solvers are more limited since we are optimizing the location of barycenters. However, thanks to the celebrated properties (Brenier, 1991; McCann, 1997), which connects the Monge pushforward map with the coupling, we can avoid the intractable optimization over μ_t and turn to the training distribution with a transport map.

Interpolation with OT map When $\nu_0, \nu_1 \in \mathcal{M}(\mathcal{X})$, ν_1 is absolutely continuous with respect to ν_0. Let $\mu(t) = \{\mu_t\}_{0\le t\le 1}$ be the geodesic of displacement interpolation, then we have a unique and deterministic interpolation $\mu_t = (T_t)\#P^0$, where $T_t(x) := (1-t)x + tT(x)$ and convex function $T(\cdot)$ is the optimal transport map (Villani, 2009). Thus, the following proposition yields an objective for finding the worst-case barycenter:

Proposition 3.2. *Consider a factored map* $T^{xy} := (T^x, T^y) : \mathbb{R}^d \times \mathbb{R}^y \mapsto \mathbb{R}^d \times \mathbb{R}^y$ *defined as* $T(\cdot,*) := (T^x(\cdot), T^y(*))$. *Consequently, rewriting Eq,* (6) *with this above map yields*

$$\begin{aligned} \tilde{\mu}_{adv}^{\{P_0,P_1\}} &:= \max_{t\in[0,1]} \mathbb{E}_{x,y\sim P^0}\left[l(f_\theta(T_t^x(x)), T_t^y(y))\right] \\ \textit{where, } T_t^x(x) &= (1-t)x + tT^x(x), \\ T_t^y(y) &= (1-t)y + tT^y(y) \end{aligned} \tag{7}$$

such that the map T^{xy} *satisfies the Monge problem* $\min_{T^{xy}\#P^0=P^1} \mathbb{E}_{x,y\sim P^0}[\mathcal{D}(\{x,y\}, T^{xy}(\{x,y\}))^2]$. *Here* $\mathcal{D}(\{x_1,y_1\},\{x_2,y_2\}) = d(x_1-y_1) + L(y_1,y_2)$ *is a measure that combines the distance between samples the discrepancy between* y_1 *and* y_2 *(Courty et al., 2017b). We note here this joint cost is separable and the analysis of the generic joint cost function is left for future studies (Fan et al., 2021; Korotin et al., 2022b).*

The benefits of this formulation: (1) Rather than the barycenter, the problem uses the OT map estimation where a lot of previous works have concerned efficiency (Perrot et al., 2016), scalability (Seguy et al., 2017) and convergence (Manole et al., 2021). (2) The transport map can be estimated solely on data and stored regardless of the supervised learning task. Moreover, the map can project *out-of-sample* points, which improves the generalization. (3) For the case where real data lies on a complex manifold, it is necessary to utilize an embedding space. Specifically, we perform the computation of Wasserstein geodesic and interpolation in an embedding space that is usually more regularized (Kingma & Welling, 2013). (4) Such OT results on labeled data distributions have been shown effective (Alvarez-Melis & Fusi, 2020; 2021; Fan & Alvarez-Melis, 2022) in transfer learning tasks while we first investigate its distributional robustness. Also, interestingly, when ignoring the expectation and defining the map as $T^x(x^0) := x^1$, $T^y(y^0) := y^1$, such interpolation coincides with the popular mixup (Zhang et al., 2018) which is simple but effective even on complicated datasets.

3.2. Implicit regularization: geodesic smoothness

Data augmentation is, in general, viewed as implicit regularization (Rebuffi et al., 2021). However, our augmentation

policy also motivates an explicit regularization term to enhance the training of deep neural networks.

For a given predictive function $f_\theta^* : \mathbb{R}^d \mapsto \mathbb{R}^y$ and a loss criterion $l(\cdot,\cdot) : \mathcal{Y}\times\mathcal{Y} \mapsto \mathbb{R}$, now we consider the following *performance geodesic*,

$$\mathrm{R}^{\mathrm{g}}_{f_\theta^*,l}(t) := \mathbb{E}_{x,y\sim\mu(t)}\left[l(f(x),y)\right], \tag{8}$$

which quantifies the loss associated with the prediction task, where $\mu(t)$ is the displacement interpolation as stated after Eq. (6). The geodesic loss $\mathrm{R}^{\mathrm{geo}}_{f_\theta^*}$ provides us a new lens through which we can measure, interpret, and improve a predictive model's robustness from a geometric perspective. Since the geodesic is a smooth and continuous path connecting P^0, P^1, The factorized geodesic interpolation T^{xy} Eq. (6) allows us to formulate a new metric $\mathrm{Reg}_l^{T^{xy}}$ that measures the change of a classifier f_θ^*, under criteria $l(\cdot,\cdot)$ when gradually transporting from P^0 to P^1.

$$\int_0^1 \left|\frac{d}{dt}\int_{\mathcal{X}\times\mathcal{Y}} l(f_\theta(T_t^x(x),T_t^y(y))dP^0(x,y)\right| dt = \int_0^1 \left|\frac{d}{dt}\mathrm{R}^{T^{xy}}_{f_\theta,l}(t)\right| dt \tag{9}$$

where $\mathrm{R}^T_{f_\theta,l}(t) = \mathbb{E}_{x,y\sim P^0}\left[l(f_\theta^*(T_t^x(x)),T_t^y(y))\right]$ is the expected loss of f_θ^* at the location t on the geodesic interpolation. Thus, to robustify a classifier f_θ, we propose to use the following regularization, which promotes the *smoothness derivative along the geodesic*.

$$\mathrm{Reg}_l^T(f_\theta) := \int_0^1 \left|\frac{d}{dt}\mathrm{R}^T_{f_\theta,l}(t)\right| dt, \text{ (geodesic regularizer)} \tag{10}$$

Proposition 3.3 (Geodesic performance as data-adaptive regularization)**.** *Consider the following minimization*

$$\min_{f\in\mathcal{H}} \mathbb{E}_{x,y\sim P}[l(f(x),y)] + \frac{\lambda_1}{2}\mathit{Reg}_l^{\hat{T}}(f)^2 + \frac{\lambda_2}{2}\|f\|^2_{\mathcal{H}}.$$

where $\mathcal{H}$ is space of function. When the data distribution is $P := 1/2P_0 + 1/2P_1$ and $P_0 = \mathcal{N}(-\mu, I), P_1 = \mathcal{N}(\mu, T)$. The objective has the following form:

$$\min_\theta \mathcal{L}(\theta) := -\langle\theta,\mu\rangle + \frac{\lambda_1}{2}|\langle\theta,\mu\rangle|^2 + \frac{\lambda_2}{2}\|\theta\|_2^2$$

The optimal solution f corresponding to data assumption has the closed form $\theta^* = (\lambda_1\mu\otimes\mu + \lambda_2 I_m)^{-1}\mu$, where $\otimes$ refers to the outer product.

The illustrative result above indicates our geodesic regularizer is a *data-depend regularization adjusted by the underlying data distributions*. This agrees with the recent theoretical understanding of mixup and robustness (Zhang et al., 2021). In addition, existing robustness criteria (empirical and certified robustness) focus on the local area of individual samples within a perturbation ball $\mathcal{S} = \{\delta : \|\delta\|_\infty \le \epsilon, \epsilon > 0\}$ without considering the robustness behavior of f_θ while transitioning from $P^0_{x,y}$ to $P^1_{x,y}$. The interpolation provides a new perspective for robustness, as shown in Fig.(2).

3.3. Computation and implementation

OT map estimation Our formulation allows us to leverage off-the-shelf OT map estimation approaches. Despite recent scalable and efficient solutions that solve the Monge problem (Seguy et al., 2017), withGAN (Arjovsky et al., 2017), ICNN (Makkuva et al., 2020), flow (Huang et al., 2020; Makkuva et al., 2020; Fan et al., 2021), and so on (Korotin et al., 2022b; Bunne et al., 2022). As the first step, in this work, we refer to barycentric projection (Ambrosio et al., 2005) thanks to its flexibility (Perrot et al., 2016) and recent convergence understandings (Manole et al., 2021; Pooladian & Niles-Weed, 2021). Specifically, we are not given access to the probability distributions P^0 and P^1, but only samples $\{(x_i^0, y_i^0)\}_{i=1}^n \sim P^0_{xy}$ and $\{(x_i^1, y_i^1)\}_{i=1}^n \sim P^1_{xy}$. When an optimal map $T_0 \in \{T : \mathcal{X}\times\mathcal{Y} \mapsto \mathcal{X}\times\mathcal{Y} | T\#P^0 = P^1\}$ exists, we want an estimator $\hat{T}_n$ with empirical distributions $\hat{P}^0_{xy,n}$ and $\hat{P}^1_{xy,n}$. This problem is usually regarded as the two-sample estimation problem (Manole et al., 2021). Among them, the celebrated Sinkhorn algorithm (Cuturi, 2013) solves the entropic Kantorovich objective efficiently $\hat{\pi}\epsilon, n := \arg\min_{\pi\in\Pi(\mathbb{R}^{n\times n})}\langle\pi, M\rangle + \epsilon H(\pi)$, where $H(\pi) := -\sum_{ij}\pi_{ij}\log\pi_{ij}$ is the negative entropy, provides an efficient $O(n^2)$ solution. Furthermore, It was shown in Section 6.2 of (Pooladian & Niles-Weed, 2021) that an entropy-regularised map (given by $\hat{T}_{\epsilon,n}$) for empirical distributions can approximate the population transportation map efficiently.

From the computation perspective, $T_{\epsilon,n}$ can be realized through the entropically optimal coupling (Cuturi, 2013) π_e^* as $\hat{T}_{\epsilon,n}(X_0) = \mathrm{diag}(\pi^*\mathbf{1}_{n_1})^{-1}\pi_\epsilon^* X_1$. Since we assume the samples are drawn i.i.d, then $\hat{T}_{\epsilon,n}(X_0) = n_0\pi^* X_1$. At this point, we can have the geodesic interpolation computed via regularized OT as

$$\hat{T}_t^{xy} := \begin{cases}\hat{T}_t^x(X_0) &= (1-t)X_0 + t\hat{T}_{\epsilon,n}(X_0),\\ \hat{T}_t^y(Y_0) &= (1-t)Y_0 + tY_1\end{cases} \tag{11}$$

Since the labels Y^i have a Dirac distribution within each P_i, we utilize the map $T^y(y^0) := y^1$. With such closed-form interpolation, we then generate samples by optimizing Eq. (7) using numerical methods such as grid search. Notice that, when setting $T^x(X_0) := X_1$ then the geodesic turns into the mixup however it can not deal with the case where we have different numbers of samples $n_0 \neq n_1$.

Regularization With the aforementioned parameterization, we thus have the following expression for the smoothness regularizer $\mathrm{Reg}_l^T(f_\theta)$.

$$\begin{aligned}\mathrm{Reg}_l^{\hat{T}}(f_\theta) &= \int_0^1 \left|\frac{d}{dt}\mathrm{R}^{T^{xy}}_{f_\theta,l}(t)\right| dt \text{ where } \frac{d}{dt}\mathrm{R}^{T^{xy}}_{f_\theta,l}(t)\\ &= \int \frac{dL}{d\tilde{y}}\nabla f_\theta((1-t)x + t\hat{T}_{\epsilon,n}(x))(\hat{T}_{\epsilon,n}(x) - x)dP^0(x,y)\end{aligned} \tag{12}$$

where $\tilde{y} = f_\theta((1-t)x + t\hat{T}(x))$ is the predicted target, $L = l(\tilde{y}, \hat{T}_t^y(y))$ measures the loss with regard to the interpolated target. This regularizer can be easily computed with the Jacobian of f_θ on mixup samples. In fact, we regularize the inner product of the expected loss, the Jacobian on interpolating samples, and the difference of $\hat{T}_{\epsilon,n}(x_0) - x_0$. In fact, Jacobian regularizer has already been shown to improve adversarial robustness (Hoffman et al., 2019)

Manifold and feature space The proposed data augmentation and regularization paradigms using the OT formulation have nice properties when data lie in a Euclidean space. However, the real data may lie on a complex manifold. In such a scenario, we can use an embedding network $\phi : \mathcal{X} \mapsto \mathcal{Z}$ that projects it to a low-dimensional Euclidean space $Z \in \mathbb{R}^z$ and a decoder network $\psi : \mathcal{Z} \mapsto \mathcal{X}$. Similarly, data augmentation may be carried out via

$$\tilde{\mu}_{adv}^{\{P_0,P_1\}} = \max_{t\in[0,1]} \mathbb{E}_{x,y\sim P^0}\left[l(f_\theta(\psi(T_t^x(\phi(x)))), T_t^y(y))\right]. \tag{13}$$

Also, regularize the geodesic on the manifold

$$\mathrm{Reg}_l^{\hat{T}}(f_\theta) = \int_0^1 \left|\frac{d}{dt}\mathrm{R}_{f_\theta,l}^{T^{xy}}(t)\right| dt, \text{ where } \frac{d}{dt}\mathrm{R}_{f_\theta,l}^{T^{xy}}(t)$$
$$= \frac{1}{n}\frac{dL}{d\tilde{y}}\nabla f_\theta((1-t)\phi(X_0) + tn\pi_\epsilon^*\phi(X_1))(n\pi_\epsilon^*\phi(X_1) - \phi(X_0)). \tag{14}$$

In fact, such a method has shown effectiveness in Manifold Mixup (Verma et al., 2019) since a well-trained embedding helps create semantically meaningful samples than mixing pixels, even on a Euclidean distance that mimics Wasserstein distance (Courty et al., 2017a). Recent advances in representation learning (Radford et al., 2021) also facilitate this idea. In this work, we also adopt Variational Autoencoder (VAE) (Kingma & Welling, 2013) and pre-trained ResNets to obtain the embedding space. We follow a standard adversarial training scheme where we iteratively (a) update the predictor f_θ with objective $\min_\theta \mathbb{E}_{x_i,y_i\sim\sum P^i+\sum \nu_{adv}[l(f_\theta(x),y)]} + \lambda\mathrm{Reg}(f_\theta)$, where the geodesic regularization $\mathrm{Reg}(f_\theta)$ is computed via Eq.(14). (b) Find and store the augmented data by maximizing the equation (6). A pseudocode Algo.(2) is attached in the Appendix.

Scalability and batch OT To handle large-scale datasets, we follow the concept of minibatch optimal transport (Fatras et al., 2021b; 2019; Nguyen et al., 2022) where we sample a batch of b_n samples from only two classes during the data augmentation procedure. Whereas minibatch OT could lead to non-optimal couplings, it contributes to better domain adaptation performance (Fatras et al., 2021a) without being concerned by the limitations of algorithmic complexity. Further, our experimental results have demonstrated that our data augmentation is still satisfactory.

Algorithm 1 The data augmentation algorithm

for a batch of mixture data **do**
  We only sample from two classes. $\{X_{s,i}, y_{s,i}\}_{i=1}^{n_s}$, and $\{X_{t,j}, y_{t,j}\}_{j=1}^{n_t}$.
  Map to manifold $Z_s = \phi_{rep}(X_s)$, $Z_t = \phi_{rep}(X_t)$.
  # Obtain the worst-case Wasserstein barycenters
  Empirical distributions $\hat{\nu}_s = \sum_{i=1}^{n_s} p_{s,i}\delta_{Z_s}$, $\hat{\nu}_t = \sum_{j=1}^{n_t} p_{t,j}\delta_{Z_y}$.
  Sample a list of $\{t_i\} \sim U[0,1]$ for Monte-Carlo
  for each t_i **do**
    # Get transported samples via (11)**.**
    Get the pushforward measures on the geodesic $\mu_{t_i} = (T_{t_i})\#\hat{\nu}_s$,
    Map to data space $\hat{X} = \phi^{-1}(\hat{Z}_s)$, $\hat{Z}_s \sim \mu_{t_i}$.
    Prediction on the geodesic $\tilde{y} = f(\hat{T}_{t_i}^x(\hat{X}))$, Compute the loss $L = l(\tilde{y}, \hat{T}_t^y(y))$.
  end for
  Store the worst-case data $\hat{X}_{t_i}$ of this batch.
end for
return All the augmented data

4. Theoretical analysis

Here, we rigorously investigate how the interpolation along geodesics affects the decision boundary and robustness. We use the concrete and natural Gaussian model (Schmidt et al., 2018) since it is theoretically tractable and, at the same time, reflects underlying properties in high-dimensional learning problems. In fact, such settings have been widely studied to support theoretical understandings in complex ML problems such as adversarial learning (Carmon et al., 2019; Dan et al., 2020), self-supervised learning (Deng et al., 2021), and neural network calibration (Zhang et al., 2022). More importantly, recent advances in deep generative models such as GAN (Goodfellow et al., 2020), VAE (Kingma & Welling, 2013), and diffusion model (Song & Ermon, 2019) endorse the theoretical analysis of Gaussian generative models (Zhang et al., 2021) on complicated real-world data, akin to our manifold geodesic setting.

Problem set-up As introduced in our preliminary (Sec. 2) and problem formulation (Sec. 3), we consider a $2-$ classification task where all the data $S = \{(x_1, y_1), ..., (x_n, y_n)\} \sim \mathcal{P}_{x,y}^{all}$ are sampled from a joint distribution $\mathcal{P}_{x,y}^{all} = 1/2P_{xy}^0 + 1/2P_{xy}^1$. Formally,

Definition 4.1 (Conditional Gaussian model)**.** For $\mu \in \mathbb{R}^d$ and $\sigma \geq 0$, the model is defined as the distribution over $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} := \{-1, 1\}$,

$$(x, y) \sim P_{xy}^{all} := p(y = -1)P_{xy}^0 + p(y = 1)P_{xy}^1$$

where $P_{xy}^0 = N(-\mu, \sigma^2 I)$, $P_{xy}^1 = N(\mu, \sigma^2 I)$, and $p(y = 1) = p(y = -1) = \frac{1}{2}$.

The goal here is is to learn a classifier $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ parameterized by θ which minimizes population classification

error

$$R_{\mu,\sigma}(f) := \mathbb{E}_{x,y\sim P_{xy}^{all}}[\mathbb{I}(f_\theta(x) \neq y)]. \tag{15}$$

And such classifier is estimated from n_0 observed training data samples $\{x_i, y_i\}_{i=1}^{n_0} \sim P_{x,y}^{all}$.

4.1. Geodesic between Gaussian distributions

The assumption of Gaussian distribution not only provides a tractable guarantee for classification but also allows us to employ desirable conclusions for optimal transport between Gaussian distributions, thanks to established studies (Dowson & Landau, 1982; Givens & Shortt, 1984; Knott & Smith, 1984). More specifically, although the Wasserstein distance and the transport map between regular measures rarely admit closed-form expressions, the W_2-Wasserstein distance between Gaussian measures can be obtained explicitly (Dowson & Landau, 1982). Moreover, the Optimal transport map (Knott & Smith, 1984) between two Gaussian distributions as well as the constant speed geodesics between them, termed McCann's interpolation have explicit forms (McCann, 1997). Please see the Appendix for more detail on this.

Given explicit forms of the augmented distribution, to proceed with the theoretical analysis, we construct the following data augmentation scheme: We always consider a symmetric pair of augmented Gaussian distributions G_t^g and G_{1-t}^g, $t \in [0, \frac{1}{2})$. Further, since the marginal distributions of the Gaussian model are $P^1 = N(-\mu, \sigma^2 I)$ and $P_2 = N(+\mu, \sigma^2 I)$, we can denote the augmented data distribution as a mixture of Gaussian distributions as $P_{xy}^{A(t)} := \frac{1}{2}G_t^g + \frac{1}{2}G_{1-t}^g$ with augmented samples $\tilde{x}_i, \tilde{y}_i : x|y \sim N(ry\mu, \sigma^2 I), y \sim U\{-1, 1\}$, where $r = 1 - 2t$.

4.2. Data augmentation improves robustness, provably

Here, we demonstrate that the data augmentation obtained via Wasserstein geodesic perturbation provably improves the robust accuracy probability. In this work, we consider the ℓ_p norm ball with $p = \infty$. Specifically, we apply a bounded worst-case perturbation before feeding the sample to the classifier. Then, we recap the definition of the standard classification and robust classification error.

Definition 4.2 (Standard, robust, and smoothed classification error probability (Schmidt et al., 2018; Carmon et al., 2019))**.** Let the original data distribution be $P_{X,Y}$. The standard error probability for classifier $f_\theta : \mathbb{R}^d \mapsto \mathcal{Y}$ is defined is $\mathbb{PE}_a(f_\theta) = P_{(x,y)\sim P_{X,Y}}(f_\theta(x) \neq y)$. The robust classification error probability is defined as $\mathbb{PE}_a^{p,\epsilon}(f_\theta) = P_{(x,y)\sim P_{X,Y}}(\exists u' \in \mathcal{B}_\epsilon^p(x), f_\theta(u') \neq y)$, where the perturbations in a ℓ_p norm ball $\mathcal{B}_\epsilon^p(x) := \{x' \in \mathcal{X} | \|x' - x\| \leq \epsilon\}$ of radius ϵ around the input. The smoothed classification error for certifiable robustness is defined as $\mathbb{PE}_a^{\sigma_s}(f_\theta) = P_{(x,y)\sim P_{X,Y},\delta\sim N(0,\sigma_s^2 I)}(f_\theta(x + \delta) \neq y)$.

Remark: In this setting, standard, robust, and smoothed accuracies are aligned (Carmon et al., 2019). In other words, a highly robust standard classifier will (1) also be robust to ℓ_∞ perturbation and (2) will be a robust base classifier for certifiable robustness, as randomized smoothing is a *Weierstrauss transform* of the deterministic base classifier. The following result provides a bound for the robust error in such a case compared to using the original data alone.

Theorem 4.3. *Suppose the original data $\{X_i, Y_i,\}_{i=1}^{n_0}$ are iid with distribution satisfying $X|Y = y \sim N(y\mu, \sigma^2 I)$, $Y \sim U\{-1, 1\}$. Additionally our $n_1 = m_a n_0$ augmented data $\{\tilde{X}_i, \tilde{Y}_i\}_{i=1}^{n_1}$ are iid and independent of the original data and satisfies $X|Y = y \sim N(yr\mu, \sigma^2 I)$, $Y \sim U\{-1, 1\}$ where $r \in [0, 1]$, are the Wasserstein geodesic interpolation. Then for $\epsilon > 0$, for all $n_1 > N_1$ with N_1 (depending on t) satisfying $tN_1 + n_0 > \sqrt{N_1}\log(N_1)$ and $\sqrt{n_0}(r + n_0)/(\sqrt{N_1}\sigma) \leq \log N_1$, then we have*

$$\mathbb{PE}_a^{p,\epsilon}(f_{\hat{\theta}_{n_0}}) \geq \mathbb{PE}_a^{p,\epsilon}(f_{\hat{\theta}_{n_0,n_1}}), \tag{16}$$

$$w.p. \geq P_{X\sim N(0,I_d)}(\|X\| \leq \log(n_1)/\sigma) \times P_{X\sim N(0,I_d)}(A \cap B), \tag{17}$$

where $A = \{X : \sqrt{n_0}(r + n_0)/(\sqrt{n_1}\sigma) \leq \|X\| \leq \log n_1\}$, $B = \{X : \mu^T X/\|X\| \leq \sqrt{\|\mu\|^2 - \{n_0(r + n_0)^2/(n_1\sigma^2)\}}$, and $\hat{\theta}_{n_0,n_1} = (\sum_{i=1}^{n_0} Y_i X_i + \sum_{i=1}^{n_1} \tilde{Y}_i \tilde{X}_i)/n_0 + n_1$, $\hat{\theta}_{n_0} = (\sum_{i=1}^{n_0} Y_i X_i)/n_0$. $f_\theta(x)$ defined as $sign(x^T\theta)$.

Note that in the above theorem $P_{X\sim N(0,I_d)}(\|X\| \leq \log(n_1)/\sigma) \to 1$ as $n_1 \to \infty$. Moreover, $P(A)$ and $P(B) \to 1$ as well with $n \to \infty$. The proof is shown in the Appendix.

Remark: From the above theory, we can see that the robustness from data augmentation will increase when we have a larger r and sample size. This explains the property of interpolation-based methods where the samples from vicinal distributions are desired. Actually, the effect of sample size can be observed in the following experiments (figure (3)).

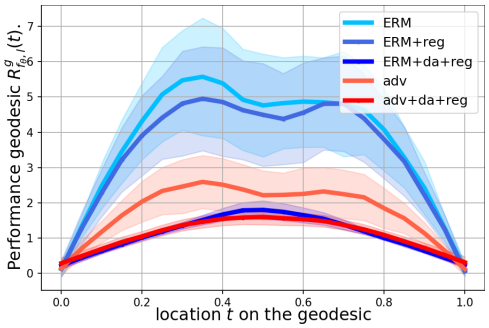


Figure 2: The *performance geodesic* $\mathrm{R}^{\mathrm{g}}_{f,l}(t)$ of classifiers obtained from different training strategies. Here *ERM* stands for normal training, *+reg* is using geodesic regularization, *+da* means using data augmentation, *adv* denotes adversarial training with PGD. Their robustness acc. are in table.(1)

5. Experiments and Discussion

We evaluate our proposed method in terms of both empirical robustness and certified robustness (Cohen et al., 2019a) on the MNIST (LeCun et al., 1998), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) dataset, and samples from ImageNet(64×64) dataset (Deng et al., 2009; Le & Yang, 2015). Typically, we use data augmentation to double the size of the training set $m_a = 1$ at first and use the regularization with a fixed weight $\alpha_{reg} = 5.0$ during the training, as implicit data augmentation. For the MNIST samples, we train a LeNet (LeCun et al., 1998) classifier with a learning rate $lr = 0.01$ for 90 epochs. For the CIFAR dataset, we use the ResNet110 (He et al., 2016) for the certifiable task on CIFAR10 and PreactResNet18 on CIFAR-100. The Sinkhorn entropic coefficient is chosen to be $\epsilon = 0.01$. We use VAE with different latent dimensions as embedding, where details can be found in the appendix. The experiments are carried out on several NVIDIA RTX A6000 GPUs and two NVIDIA GeForce RTX 3090 GPUs.

Table 1: Adversarial robustness (MNIST), ours:DA + Reg

	ERM	PGD	ERM+**Reg**	ERM + **Ours**	PGD +**Ours**
Clean Acc.(%)	99.09 ± 0.02	98.94 ± 0.03	99.27 ± 0.06	$\mathbf{99.39 \pm 0.03}$	99.34 ± 0.02
Robust Acc.(%)	31.47 ± 0.12	81.23 ± 0.17	35.66 ± 0.20	81.23 ± 0.12	$\mathbf{82.74 \pm 0.16}$

Empirical robustness We use the strongest PGD method to perform ℓ_∞ attack with $\epsilon = 2.0$ and 4 steps. As shown in table (1), either data augmentation and regularization can improve the ℓ_∞ robustness. Moreover, our method can further improve gradient-based adversarial training. In fig.(2), we visualize the *performance geodesic* of various training strategies where more robust models apparently have mode smoother geodesic. For empirical robustness on CIFAT-100 and ImageNet(64×64), we follow training protocol from (Kim et al., 2020) and compare our method with ERM, vanilla mixup (Zhang et al., 2018), Manifold mixup (Verma et al., 2019), CutMix (Yun et al., 2019), AugMix (Hendrycks et al., 2019), and PuzzleMix (Kim et al., 2020). To enable a fair comparison, we reproduce only the non-adversarial PuzzleMix methods.

For ImageNet(64×64), it contains 200 classes with 64×64 resolution(Chrabaszcz et al., 2017). As in table (2), our approach outperforms the best baseline by 16.8% under ($\epsilon = 4/255$) FGSM attack. For CIFAR-100, as in table (3), we exceed baseline by 22.4% under FGSM, while having comparable performance in standard accuracy. It may be caused by our data augmentation populating samples only on the local decision boundaries.

Table 2: ℓ_∞ Standard and adversarial robust accuracy on ImageNet (64×64) with baselines

	ERM	Mixup	Manifold	CutMix	AugMix	PuzzleMix	**Ours**
Clean Acc.(%)	58.28	56.32	58.08	57.71	56.135	63.31	**64.15**
FGSM Acc.(%)	8.22	11.21	10.69	11.61	11.07	7.82	**13.46**

Table 3: ℓ_∞ Standard and adversarial robust accuracy on CIFAR-100 dataset with baselines

	ERM	Mixup	Manifold	CutMix	AugMix	PuzzleMix	**Ours**
Clean Acc.(%)	78.76	81.90	81.98	82.31	79.82	83.05	81.36
FGSM Acc.(%)	34.72	43.21	39.22	30.81	44.82	36.82	**54.87**

ℓ_2 Certifiable robustness We compare the performance with Gaussian (Cohen et al., 2019b), Stability training (Li et al., 2018), SmoothAdv (Salman et al., 2019), MACER (Zhai et al., 2020), Consistency (Jeong & Shin, 2020), and SmoothMix (Jeong et al., 2021). We follow the same evaluation protocol proposed by (Cohen et al., 2019b) and used by previous works (Salman et al., 2019; Zhai et al., 2020; Jeong & Shin, 2020), which is a Monte Carlo-based certification procedure, calculating the prediction and a "safe" lower bound of radius over the randomness of n samples with probability at least $1 - \alpha$, or abstaining the certification. We consider three different models as varying the noise level $\sigma \in \{0.25, 0.5, 1.0\}$. During inference, we apply randomized smoothing with the same σ used in training. The parameters in the evaluation protocol (Cohen et al., 2019b) are set as: $n = 100,000$, $n_0 = 100$, and $\alpha = 0.001$, with previous work (Cohen et al., 2019b; Jeong & Shin, 2020).

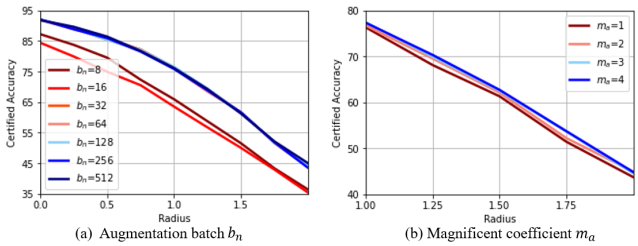
Table 4: Certified accuracy on CIFAR-10 dataset.

σ	Models (CIFAR-10)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25
0.25	Gaussian	76.6	61.2	42.2	25.1	0.0	0.0	0.0	0.0	0.0	0.0
	Stability training	72.3	58.0	43.3	27.3	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv*	73.4	65.6	57.0	47.1	0.0	0.0	0.0	0.0	0.0	0.0
	MACER*	79.5	69.0	55.8	40.6	0.0	0.0	0.0	0.0	0.0	0.0
	Consistency	75.8	67.6	58.1	46.7	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothMix	77.1	67.9	57.9	46.7	0.0	0.0	0.0	0.0	0.0	0.0
	Ours	69.7	**63.4**	**56.0**	**47.3**	0.0	0.0	0.0	0.0	0.0	0.0
0.50	Gaussian	65.7	54.9	42.8	32.5	22.0	14.1	8.3	3.9	0.0	0.0
	Stability training	60.6	51.5	41.4	32.5	23.9	15.3	9.6	5.0	0.0	0.0
	SmoothAdv*	65.3	57.8	49.9	41.7	33.7	26.0	19.5	12.9	0.0	0.0
	MACER*	64.2	57.5	49.9	42.3	34.8	27.6	20.2	12.6	0.0	0.0
	Consistency	64.3	57.5	50.6	43.2	36.2	29.5	22.8	16.1	0.0	0.0
	SmoothMix	65.0	56.7	49.2	41.2	34.5	29.6	23.5	18.1	0.0	0.0
	Ours	55.4	50.7	**45.6**	**40.6**	**35.2**	**30.5**	**25.7**	**19.5**	0.0	0.0
1.00	Gaussian	47.2	39.2	34.0	27.8	21.6	17.4	14.0	11.8	10.0	7.6
	Stability training	43.5	38.9	32.8	27.0	23.1	19.1	15.4	11.3	7.8	5.7
	SmoothAdv*	50.8	44.9	39.0	33.6	28.1	23.7	19.4	15.4	12.0	8.7
	MACER*	40.4	37.5	34.2	31.3	27.5	23.4	22.4	19.2	16.4	13.5
	Consistency	46.3	41.8	37.9	34.2	30.1	26.1	22.3	19.7	16.4	13.8
	SmoothMix	47.1	42.5	37.5	32.9	28.2	24.9	21.3	18.3	15.5	12.6
	Ours	40.9	36.9	**34.5**	**31.5**	**28.1**	**24.4**	**22.5**	**20.4**	**16.5**	**14.0**

For the CIFAR-10 dataset, Table 4 showed that our method generally exhibited better certified robustness compared to other baselines, i.e., SmoothAdv (Salman et al., 2019), MACER (Zhai et al., 2020), Consistency (Jeong & Shin, 2020), and SmmothMix (Jeong et al., 2021). The important characteristic of our method is the robustness under larger noise levels. Our method achieved the highest certified test accuracy among all the noise levels when the radii are large, i.e., radii 0.25-0.75 under noise level $\sigma = 0.25$, radii 0.50-1.75 under noise level $\sigma = 0.50$, and radii 0.50-2.75 under noise level $\sigma = 1.00$. We obtained the new state-of-the-art certified robustness performance under large radii. As shown in Table 4, we find that by combing our data augmentation mechanism, the performance of previous SOTA methods can be even better, which demonstrates our

method can be easily used as an add-on mechanism for other algorithms to improve robustness.

Ablation study Here, we conducted detailed ablation studies to investigate the effectiveness of each component. We performed experiments on MNIST with $\sigma = 1.0$, and the results are shown in Figure 3. (More results are shown in the Appendix.) We investigated the influence by the augmentation batch size b_n (Figure 3(a)) following batch optimal transport setting. Also, we studied data augmentation sample size m_a (Figure 3(b)). We found that a larger augmentation batch size b_n leads to better performance, which is expected since it better approximates the joint measure. In addition, more augmented data samples m_a benefit robustness, which agrees with our theoretical results.



(a) Augmentation batch b_n (b) Magnificent coefficient m_a

Figure 3: Comparison of approximate certified accuracy with different parameters, including (a) augmentation batch size b_n, and (b) magnificent coefficient m_a.

6. Conclusion and Future Work

In this paper, we proposed to characterize the robustness of a model through its performance on the Wasserstein geodesic connecting different training data distributions. The worst-case distributions on the geodesic allow us to create augmented data for training a more robust model. Further, we can regularize the smoothness of a classifier to promote robustness. We provided theoretical guarantees and carried out extensive experimental studies on multiple datasets including CIFAR100 and ImageNet. Our results showed new SOTA performance compared with existing methods and can be easily combined with other learning schemes to boost the robustness.

As a first step, this work provides a new perspective to characterize the model's robustness. There could be several future works, including considering multi-marginal adversarial Wasserstein barycenter on a simplex, more efficient optimization on the geodesic, and more thorough theoretical studies beyond Gaussian models. Additionally, other notions of Wasserstein distances (Guha et al., 2023) could also be used to extend the current set of results under specific objectives.

Acknowledgements

This work is partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR00112320012. We would like to thank Kilian Fatras, Lin Gui, and anonymous reviewers for their valuable feedback and comments on our manuscript.

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Alvarez-Melis, D. and Fusi, N. Dataset dynamics via gradient flows in probability space. In *International Conference on Machine Learning*, pp. 219–230. PMLR, 2021.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Amos, B. On amortizing convex conjugates for optimal transport. *arXiv preprint arXiv:2210.12153*, 2022.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. *arXiv preprint arXiv:1909.12272*, 2019.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional monge maps. *arXiv preprint arXiv:2206.14262*, 2022.

Cao, C., Zhou, F., Dai, Y., and Wang, J. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *arXiv preprint arXiv:2212.10888*, 2022.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019a.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019b.

Courty, N., Flamary, R., and Ducoffe, M. Learning wasserstein embeddings. *arXiv preprint arXiv:1710.07457*, 2017a.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017b.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, Z., Zhang, L., Ghorbani, A., and Zou, J. Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*, pp. 2845–2853. PMLR, 2021.

Dong, Y., Deng, Z., Pang, T., Zhu, J., and Su, H. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33: 8270–8283, 2020.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Fan, J. and Alvarez-Melis, D. Generating synthetic datasets by interpolating along generalized geodesics. In *NeurIPS*

2022 Workshop on Synthetic Data for Empowering ML Research, 2022.

Fan, J., Taghvaei, A., and Chen, Y. Scalable computations of wasserstein barycenter via input convex neural networks. *arXiv preprint arXiv:2007.04462*, 2020.

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H. Scalable computation of monge maps with general costs. *arXiv preprint arXiv:2106.03812*, pp. 4, 2021.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Fatras, K., Séjourné, T., Flamary, R., and Courty, N. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pp. 3186–3197. PMLR, 2021a.

Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021b.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Gao, Y. and Chaudhari, P. An information-geometric distance on the space of tasks. In *International Conference on Machine Learning*, pp. 3553–3563. PMLR, 2021.

Givens, C. R. and Shortt, R. M. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Greenewald, K., Gu, A., Yurochkin, M., Solomon, J., and Chien, E. k-mixup regularization for deep learning via optimal transport. *arXiv preprint arXiv:2106.02933*, 2021.

Guha, A., Ho, N., and Nguyen., X. On excess mass behavior in gaussian mixture models with orlicz-wasserstein distances. 2023.

He, K. and Sun, J. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5353–5360, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Hoffman, J., Roberts, D. A., and Yaida, S. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Howard, A. G. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

Hua, X., Nguyen, T., Le, T., Blanchet, J., and Nguyen, V. A. Dynamic flows on curved space generated by labeled data. *arXiv preprint arXiv:2302.00061*, 2023.

Huang, C.-W., Chen, R. T., Tsirigotis, C., and Courville, A. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. *arXiv preprint arXiv:2012.05942*, 2020.

Huang, P., Xu, M., Zhu, J., Shi, L., Fang, F., and Zhao, D. Curriculum reinforcement learning using optimal transport via gradual domain adaptation. *arXiv preprint arXiv:2210.10195*, 2022.

Janati, H., Cuturi, M., and Gramfort, A. Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, pp. 4692–4701. PMLR, 2020.

Jeong, J. and Shin, J. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems*, 2020.

Jeong, J., Park, S., Kim, M., Lee, H.-C., Kim, D.-G., and Shin, J. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Kim, H., Lee, W., Lee, S., and Lee, J. Bridged adversarial training. *arXiv preprint arXiv:2108.11135*, 2021a.

Kim, J.-H., Choo, W., and Song, H. O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pp. 5275–5285. PMLR, 2020.

Kim, J.-H., Choo, W., Jeong, H., and Song, H. O. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021b.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Knott, M. and Smith, C. S. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

Korotin, A., Egiazarian, V., Li, L., and Burnaev, E. Wasserstein iterative networks for barycenter estimation. *arXiv preprint arXiv:2201.12245*, 2022a.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022b.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998.

Lewy, D. and Mańdziuk, J. An overview of mixing augmentation methods and augmentation strategies. *Artificial Intelligence Review*, pp. 1–59, 2022.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. *arXiv preprint arXiv:1809.03113*, 2018.

Li, L., Weber, M., Xu, X., Rimanic, L., Kailkhura, B., Xie, T., Zhang, C., and Li, B. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 535–557, 2021.

Lin, T., Ho, N., Cuturi, M., and Jordan, M. I. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(65): 1–43, 2022.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113. PMLR, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Mroueh, Y. and Nguyen, T. On the convergence of gradient descent in gans: Mmd gan as a gradient flow. In *International Conference on Artificial Intelligence and Statistics*, pp. 1720–1728. PMLR, 2021.

Ng, N., Cho, K., and Ghassemi, M. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *ArXiv*, abs/2009.10195, 2020.

Nguyen, K., Nguyen, D., Pham, T., Ho, N., et al. Improving mini-batch optimal transport via partial transportation. In *International Conference on Machine Learning*, pp. 16656–16690. PMLR, 2022.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Qiu, J., Zhu, J., Rosenberg, M., Liu, E., and Zhao, D. Optimal transport based data augmentation for heart disease diagnosis and prediction. *arXiv preprint arXiv:2202.00567*, 2022a.

Qiu, J., Zhu, J., Xu, M., Dernoncourt, F., Bui, T., Wang, Z., Li, B., Zhao, D., and Jin, H. Semantics-consistent cross-domain summarization via optimal transport alignment. *arXiv preprint arXiv:2210.04722*, 2022b.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark,

J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

Tokozume, Y., Ushiku, Y., and Harada, T. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*, 2017.

Uddin, A., Monira, M., Shin, W., Chung, T., Bae, S.-H., et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

Venkataramanan, S., Kijak, E., laurent amsaleg, and Avrithis, Y. Alignmix: Improving representations by interpolating aligned features, 2022. URL `https://openreview.net/forum?id=jFlWZEv6dv`.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Wang, H., Li, B., and Zhao, H. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, pp. 22784–22801. PMLR, 2022.

Weber, M., Li, L., Wang, B., Zhao, Z., Li, B., and Zhang, C. Certifying out-of-domain generalization for blackbox functions. *arXiv preprint arXiv:2202.01679*, 2022.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Xiao, C., Zhong, P., and Zheng, C. Enhancing adversarial defense by k-winners-take-all. *arXiv preprint arXiv:1905.10510*, 2019.

Xing, Y., Song, Q., and Cheng, G. Unlabeled data help: Minimax analysis and adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 136–168. PMLR, 2022.

Yao, H., Zhang, L., and Finn, C. Meta-learning with fewer tasks through task interpolation. *arXiv preprint arXiv:2106.02695*, 2021.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C.-J., and Wang, L. MACER: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Zhang, H. and Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8yKEo06dKNo.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. In *International Conference on Machine Learning*, pp. 26135–26160. PMLR, 2022.

Zhu, J., Shi, L., Yan, J., and Zha, H. Automix: Mixup networks for sample interpolation via cooperative barycenter learning. In *European Conference on Computer Vision*, pp. 633–649. Springer, 2020.

Zhu, J., Guha, A., Do, D., Xu, M., Nguyen, X., and Zhao, D. Functional optimal transport: map estimation and domain adaptation for functional data, 2021.

Zhu, J., Darnell, G., Kumar, A., Zhao, D., Li, B., Nguyen, X., and Ren, S. Y. Physiomtl: Personalizing physiological patterns using optimal transport multi-task regression. *arXiv preprint arXiv:2203.12595*, 2022a.

Zhu, J., Qiu, J., Yang, Z., Weber, D., Rosenberg, M. A., Liu, E., Li, B., and Zhao, D. Geoecg: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction. *arXiv preprint arXiv:2208.01220*, 2022b.

Appendix

Table of Contents

A Theoretical results **19**
- A.1 The Wasserstein distance between Gaussian distribution . . . 19
- A.2 Transport map between Gaussian measures . . . 20
- A.3 The effect of the regularization . . . 21
- A.4 Proof of Proposition 3.1 . . . 22
- A.5 Error probabilities in closed form (Definition 4.2) . . . 24
- A.6 Provable improvement from data augmentation (Theorem 4.3) . . . 25

B Additional intuition & related works **26**
- B.1 Related works . . . 26
- B.2 Intuitions . . . 26
- B.3 Mixture of distributions . . . 26
- B.4 Connection to Mixup . . . 27
- B.5 A more general problem formulation: multi-marginal optimal transport . . . 27

C Algorithm & Computation **28**

D Additional Experimental Results **29**
- D.1 Experiments on MNIST dataseet . . . 29
- D.2 More ablation studies . . . 30
- D.3 More experimental results . . . 31

A. Theoretical results

A.1. The Wasserstein distance between Gaussian distribution

Theorem A.1 (The Wasserstein distance between Gaussian distribution). *Given two Gaussian measures, $P^0 = N(\mu_0, \Sigma_0)$ and $P^1 = N(\mu_1, \Sigma_1)$. The L^2-Wasserstein distance between them is given by*

$$\begin{aligned} W_2^2(P^0, P^1) =& \|\mu_0 - \mu_1\|^2 + tr(\Sigma_0) + tr(\Sigma_1) \\ & - 2tr(\sqrt{\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}}) \end{aligned} \tag{18}$$

The proof largely depends on a summary of (Givens & Shortt, 1984), as well as. We recall the theorem. Given two Gaussian measures, $P^0 = N(\mu_0, \Sigma_0)$ and $P^1 = N(\mu_1, \Sigma_1)$. The L^2-Wasserstein distance between them is given by

$$W_2^2(P^0, P^1) = \|\mu_0 - \mu_1\|^2 + tr(\Sigma_0) + tr(\Sigma_1) - 2tr(\sqrt{\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}}) \tag{19}$$

Given two distributions $P^0 = N(\mu_0, \Sigma_0)$ and $P^1 = N(\mu_1, \Sigma_1)$, one can first reduce to the centered case $\mu_0 = \mu_1 = 0$. Next, let $X \sim P^0$ and $Y \sim P^1$ and if (X, Y) is a random vector of $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix

$$\Gamma = \begin{bmatrix} \Sigma_0 & C \\ C^\top & \Sigma_1 \end{bmatrix}$$

then the quantity

$$\mathbb{E}[\|X - Y\|_2^2] = Tr(\Sigma_0 + \Sigma_1 - 2C)$$

depends only on Γ. Also, when $P^0 = N(0, \Sigma_0)$ and $P^1 = N(0, \Sigma_1)$, one can restrict the infimum which defines W_2 to run over W_2 to run over Gaussian laws $N(0, \Gamma)$ on $\mathbb{R}^n \times \mathbb{R}^n$ with covariance matrix Γ structured above. Here the constrain on C is the Schur complement constraint $\Sigma_0 - C\Sigma_1^{-1}C^\top \succeq 0$. Thus, the minimization of the function $C \mapsto -2Tr(C)$ leads to the result.

A.2. Transport map between Gaussian measures

Proposition A.2 (Optimal transport map between Gaussian measures, adopted from (Knott & Smith, 1984), Example 1.7 (McCann, 1997))**.** *Given the two Gaussian distributions defined in theorem (A.1), define a symmetric positive matrix T and a corresponding linear map $\mathcal{T} : \mathcal{X} \mapsto \mathcal{X}$*

$$\begin{aligned} \mathcal{T}(x) &= \mu_1 + T(x - \mu_0), where \\ T &= \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{-1/2}\Sigma_1^{1/2} \end{aligned} \tag{20}$$

Then, $\mathcal{T}$ is the optimal map such that $P^1 = T\#P^0$.

Hence we obtain a closed-form map and that leads to an explicit form of the geodesic interpolation among two Gaussian distributions, as:

$$T_t^{G,xy} := \begin{cases} T_t^{G,x}(x^0) & = (1-t)x^0 + \mathcal{T}(x^0), \\ T_t^{G,y}(y^0) & = \begin{cases} y^0, 0 < t < 0.5, \\ -y^0, 0.5 < t < 1. \end{cases} \end{cases}$$

Here the interpolation of y^0 means we will always label a sample as -1 or 1, which allows us to proceed to the analysis under the classification error. Given the two Gaussian distributions defined in theorem (A.1), define a symmetric positive matrix T and a corresponding linear map $\mathcal{T} : \mathcal{X} \mapsto \mathcal{X}$

$$\begin{aligned} \mathcal{T}(x) &= \mu_1 + T(x - \mu_0), where \\ T &= \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{-1/2}\Sigma_1^{1/2} \end{aligned}$$

Then, $\mathcal{T}$ is the optimal map such that $P^1 = T\#P^0$.
Check the optimal transport map from $N(\mu_0, \Sigma_0)$ to $N(\mu_1, \Sigma_1)$, again assume $\mu_0 = \mu_1 = 0$ for simplicity, one may define the random column vectors $X \sim N(\mu_0, \Sigma_1)$ and $Y = TX$ and write

$$\begin{aligned} \mathbb{E}(YY^\top) &= T\mathbb{E}(XX^\top)T^\top \\ &= \Sigma_0^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_0^{1/2})^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2} \\ &= \Sigma_1 \end{aligned}$$

To show the map is optimal, one may use

$$\begin{aligned} \mathbb{E}(\|X - Y\|_2^2) &= \mathbb{E}(\|X\|_2^2) + \mathbb{E}(\|Y\|_2^2) - 2\mathbb{E}(< X, Y >) \\ &= Tr(\Sigma_0) + Tr(\Sigma_1) - 2\mathbb{E}(< X, TX >) \\ &= Tr(\Sigma_0) + Tr(\Sigma_1) - 2Tr(\Sigma_0 T) \end{aligned}$$

and one can observe that by the cyclic property of the trace

$$Tr(\Sigma_0 T) = Tr((\Sigma_1^{1/2}\Sigma_2\Sigma_0^{1/2})^{1/2})$$

Moreover, for the interpolation Geodesic, the following result holds.

Proposition A.3 (Interpolation between Gaussian distributions (McCann, 1997) example 1.7)**.** *Given the two Gaussian distributions defined in theorem (A.1), respectively. A constant speed geodesic is defined by the the path $(G_t^g)_{t\in[0,1]}$, such that $\forall t \in [0, 1]$, and $G_t^g := N(m_t, \Sigma_t)$ is given by*

$$\begin{aligned} m_t &= (1-t)\mu_1 + t\mu_2, \\ \Sigma_t &= ((1-t)Id + tT)\Sigma_1((1-t)Id + tT) \end{aligned} \tag{21}$$

where $T = \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{-1/2}\Sigma_1^{1/2}$.

We have illustrated that the Wasserstein barycenter, on the geodesic, of two Gaussian distributions is still a Gaussian distribution and admits an explicit form.

$$\begin{aligned} \text{where, } T_t^x(x) &= (1-t)x + tT^x(x), \\ T_t^y(y) &= (1-t)y + tT^y(y) \end{aligned} \tag{22}$$

A.3. The effect of the regularization

Recall that the geodesic regularization term is

$$\text{Reg}_l^{T^{xy}}(f_\theta) = \int_{t=0}^{1} \left| \frac{d}{dt} \int_{\mathcal{X}\times\mathcal{Y}} l(f_\theta(T_t^x(x), T_t^y(y)) dP^0(x,y) \right| dt \tag{23}$$

Recall the classifier to be $f_\theta(x) = \langle \theta, x \rangle = \theta^\top x$, and the loss to be $l(f(x), y) = -yf(x)$,
Consider the following minimization problem

$$\min_{f\in\mathcal{H}} \mathbb{E}_{(x,y)\sim P_{all}}[l(f(x), y)] + \frac{\lambda_1}{2}\text{Reg}_l^{T^{xy}}(f_\theta) \tag{24}$$

Proof. The problem above can be written as follows

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) = -(\langle \theta, \mu_1 - \mu_0 \rangle) + \lambda_1 \text{Reg}(\theta) \tag{25}$$

And under the above setting, we have

$$\begin{aligned}
\text{Reg}_l^{T^{xy}}(f_\theta) &= \int_{t=0}^{1} \left| \frac{d}{dt} \int_{\mathcal{X}\times\mathcal{Y}} l(f_\theta(T_t^x(x), T_t^y(y)) dP^0(x,y) \right| dt \\
&= \int_{t=0}^{1} \left| \frac{d}{dt} \mathbb{E}_{x_0.y_0\sim P^0(x,y)}[l(f_\theta(T_t^x(x), T_t^y(y))] \right| dt \\
&= \int_{t=0}^{1} \left| \frac{d}{dt} \mathbb{E}_{x_0.y_0\sim P^0(x,y)}[-T_t^y(y_0)\langle \theta, T_t^x(x_0)\rangle] \right| dt
\end{aligned} \tag{26}$$

Recall the parameterization of the map as interpolation, as e.q., 7,

Here, using the dominated convergence theorem. Also, use the closed form interpolation $T_t^x(x_0) = (1-t)x_0 + t\hat{T}(x_0)$, and $T_t^y(y_0) = 2t - 1$ ($y_0 = -1$ and $y_1 = 1$), the above equation turns into

$$= \int_{t=0}^{1} \left| \mathbb{E}_{x_0.y_0\sim P^0(x,y)}[-2\langle \theta, T_t^x(x_0)\rangle - (2t-1)\langle \theta, \hat{T}(x_0) - x_0\rangle] \right| dt \tag{27}$$

Since we have closed-form expression for the geodesic as $\mathbb{E}_{x_0\sim P^0}[T^t(x_0)] = (2t-1)\mu$ and $\mathbb{E}_{x_0\sim P^0}[\hat{T}(x_0) - x_0] = 2\mu$, then

$$= \int_{t=0}^{1} \left| -2(2t-1)\theta^\top\mu - 2(2t-1)\theta^\top\mu \right| dt = \int_{t=0}^{1} \left| 4(1-2t)\theta^\top\mu \right| dt = 2|\theta^\top\mu| \tag{28}$$

Then, the objective becomes

$$\min_\theta \mathcal{L}(\theta) := -\langle \theta, \mu \rangle + \frac{\lambda_1}{2}|\langle \theta, \mu \rangle|^2 + \frac{\lambda_2}{2}\|\theta\|_2^2 \tag{29}$$

Setting the first-order derivative to be zero, we obtain the solution

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= -\mu + \lambda_1 \mu \otimes \mu\theta + \lambda_2\theta = 0, \\
\theta^* &= (\lambda_1 \mu \otimes \mu + \lambda_2 I_m)^{-1}\mu
\end{aligned} \tag{30}$$

From Eq. (29), we can see that geodesic regularization is a data-adaptive regularization that smooths the classifier according to the distribution of original data.

A.4. Proof of Proposition 3.1

Assumptions and generalisation Without loss of generalisation, we assume $\sigma = 1$ and $\mu = (1, 0, \ldots, 0)$, with 1 at the first coordinate and 0 elsewhere. Thus $\|\mu\| = 1$.
Since $\eta(x|y=-1), \eta(x|y=1)$ are both Gaussian, assume $\eta(x|y=-1) \sim N(\psi_{-1}, \Sigma_{-1})$ and $\eta(x|y=1) \sim N(\psi_1, \Sigma_1)$.
Let $u_1, \ldots, u_d$ be the eigenvalues of Σ_0 Following Theorem A.1, we have

$$\begin{aligned}
W_2^2(N_d(\psi_1, \Sigma_1), N_d(\mu, \mathbb{I})) &= \|\psi_1 - \mu\|^2 + tr(\Sigma_1) + d - 2tr({\Sigma_1}^{1/2}) \\
&= \|\psi_1 - \mu\|^2 + \sum_{i=1}^{d} (\sqrt{u_i} - 1)^2
\end{aligned} \tag{31}$$

Similarly,

$$W_2^2(N_d(\psi_{-1},\Sigma_{-1}),N_d(-\mu,\mathbb{I})) = \|\psi_{-1}+\mu\|^2 + \sum_{i=1}^{d}\left(\sqrt{v_i}-1\right)^2, \tag{32}$$

where $\{v_i\}_1^d$ are the eigenvalues of Σ_{-1}.
By linearity property of Gaussian distributions $\eta(X|Y=y)\sim N(\psi_y,\Sigma_y)$ implies, $\alpha'X+\beta|Y=y\sim N(\alpha'\psi_y+\beta,\alpha'\Sigma_y\alpha)$. Therefore,

$$\mathbb{E}_{(X,Y)\sim\eta(X|Y)\nu(Y)}l(f_{(\alpha,\beta)}(X),Y) \propto P_{N(0,1)}\left(Z_1 < -\frac{(\alpha'\psi_1+\beta)}{\sqrt{\alpha'\Sigma_1\alpha}}\right) + P_{N(0,1)}\left(Z_{-1} > -\frac{(\alpha'\psi_{-1}+\beta)}{\sqrt{\alpha'\Sigma_{-1}\alpha}}\right), \tag{33}$$

where $Z_1 = \frac{\alpha'(X-\psi_1)}{\sqrt{\alpha'\Sigma_1\alpha}}\Big|Y=1\sim N(0,1)$ if $X|Y=1\sim N(\psi_1,\Sigma_1)$ and
$Z_{-1} = \frac{\alpha'(X-\psi_{-1})}{\sqrt{\alpha'\Sigma_{-1}\alpha}}\Big|Y=-1\sim N(0,1)$ if $X|Y=-1\sim N(\psi_{-1},\Sigma_{-1})$.
The LHS of Eq. (5) becomes

$$\max_{(\psi_1,\psi_{-1},\Sigma_1,\Sigma_{-1})\in B_\epsilon(\mu)} \inf_{\alpha,\beta} P_{N(0,1)}\left(Z_1 < -\frac{(\alpha'\psi_1+\beta)}{\sqrt{\alpha'\Sigma_1\alpha}}\right) + P_{N(0,1)}\left(Z_{-1} > -\frac{(\alpha'\psi_{-1}+\beta)}{\sqrt{\alpha'\Sigma_{-1}\alpha}}\right), \tag{34}$$

where, $B_\epsilon(\mu) = \{(\psi_1,\psi_{-1},\Sigma_1,\Sigma_{-1}) : \|\psi_{-1}+\mu\|^2 + \sum_{i=1}^d\left(\sqrt{v_i}-1\right)^2 \le \epsilon^2, \|\psi_1-\mu\|^2 + \sum_{i=1}^d\left(\sqrt{u_i}-1\right)^2 \le \epsilon^2\}$ $\{u_i\}_1^d$ are eigenvalues of Σ_{-1}, and $\{u_i\}_1^d$ are eigenvalues of Σ_{-1}.

Step 1: First, we show the proof under the assumption $\Sigma_1=\Sigma_{-1}=\Sigma$.
Fixing ψ_1,ψ_{-1},Σ so that $\psi_1,\psi_{-1},\Sigma,\Sigma\in B_\epsilon(\mu)$ and setting derivative to 0 of

$$P_{N(0,1)}\left(Z_1 < -\frac{(\alpha'\psi_1+\beta)}{\sqrt{\alpha'\Sigma\alpha}}\right) + P_{N(0,1)}\left(Z_{-1} > -\frac{(\alpha'\psi_{-1}+\beta)}{\sqrt{\alpha'\Sigma\alpha}}\right) \tag{35}$$

w.r.t β yields either (i) $\alpha'\psi_1+\beta=\alpha'\psi_{-1}+\beta$, or (ii) $\alpha'\psi_1+\beta=-(\alpha'\psi_{-1}+\beta)$.
Setting derivative of (35) w.r.t α to 0 yields $\psi_1=\psi_{-1}$, which clearly is a contradiction for $\epsilon<1$.
Considering (ii) above yields $\beta=-\alpha'(\psi_{-1}+\psi_1)/2$. Then Eq. (35) leads to

$$\begin{aligned} &P_{N(0,1)}\left(Z > \frac{\alpha'(\psi_1-\psi_{-1})}{2\sqrt{\alpha'\Sigma\alpha}}\right) + P_{N(0,1)}\left(Z < -\frac{\alpha'(\psi_1-\psi_{-1})}{2\sqrt{\alpha'\Sigma\alpha}}\right) \\ &\qquad = 2P_{N(0,1)}\left(Z > \frac{\alpha'(\psi_1-\psi_{-1})}{2\sqrt{\alpha'\Sigma\alpha}}\right). \end{aligned} \tag{36}$$

The RHS above is minimized w.r.t α when $\frac{\alpha'(\psi_1-\psi_{-1})}{2\sqrt{\alpha'\Sigma\alpha}}$ is maximized. A simple calculation shows that this is achieved for $\alpha=\frac{\tilde{a}\Sigma^{-1}}{\sqrt{\tilde{a}'\Sigma^{-1}\tilde{a}}}$, where $\tilde{a}=(\psi_1-\psi_{-1})$. This gives $\frac{\alpha'(\psi_1-\psi_{-1})}{\sqrt{\alpha'\Sigma\alpha}}=\sqrt{\tilde{a}'\Sigma^{-1}\tilde{a}}$.
For a given Σ, this is minimized (and equivalently Eq. (35) is maximised) when $(\psi_1-\psi_{-1})/2$ is in the direction of the largest eigenvector of Σ, say v_1, and $\tilde{a}'\Sigma^{-1}\tilde{a}=\|\psi_1-\psi_{-1}\|^2/v_1$.
Notice that since the other eigenvalues do not affect the choice of $(\psi_1-\psi_{-1})$, based on the constraints $\|\psi_{-1}+\mu\|^2+\sum_{i=1}^d\left(\sqrt{v_i}-1\right)^2\le\epsilon^2$ and $\|\psi_1+\mu\|^2+\sum_{i=1}^d\left(\sqrt{v_i}-1\right)^2\le\epsilon^2$, where v_i are the ordered eigenvalues of Σ, we can assume all other eigenvalues to be 1 without affecting the value in the RHS of Eq. (36).
Let the largest eigenvalue of Σ be $(1+\delta)^2$. Then based on the constraints $\|\psi_1-\psi_{-1}\|^2/v_1$ is minimized when $\psi_1=-\psi_{-1}$ and they are concurrent with μ and $-\mu$.
Based on the constraints, this yields,

$$\sqrt{\|\psi_1-\psi_{-1}\|^2/v_1} = 2(1-\sqrt{\epsilon^2-\delta^2})/(1+\delta), \tag{37}$$

which is an increasing function of δ and is minimised at $\delta = 1$, which implies $\Sigma = \mathbb{I}$, and $\psi_1 = (1-\epsilon)\mu$, $\psi_{-1} = (1-\epsilon)-\mu$, which are the geodesic optimizers.
Therefore,

$$\max_{\tilde{A}(\nu_\mu,\epsilon)} \inf_{\alpha,\beta} \mathbb{E}_{(X,Y)\sim\eta(X|Y)\nu(Y)} l(f_{(\alpha,\beta)}(X), Y) = \mathbb{E}_{(X,Y)\sim\nu_{(1-\epsilon)\mu}(X|Y)\nu(Y)} l(f_{(\mu,0)}(X), Y). \tag{38}$$

where $\tilde{A}(\nu_\mu, \epsilon) = \{\eta(X|Y) : \eta(\cdot|Y)$ is Gaussian, $W(\eta(X|Y), \nu_\mu(X|Y)) \leq \epsilon,\ \forall Y \in \{-1, 1\}$, $Cov(\eta(\cdot|Y = 1)) = Cov(\eta(\cdot|Y = -1))\}$

Step 2: Next, we use the result in Step 1 to show the remaining.
Therefore, consider the situation where we do not restrict to $\Sigma_1 = \Sigma_{-1}$.
Set $\Sigma_1^0 = \Sigma_{-1}^0 = \mathbb{I}$, $\psi_1^0 = (1-\epsilon)\mu$, $\psi_{-1}^0 = -(1-\epsilon)\mu$, $\alpha_0 = (1-\epsilon)\mu$, $\beta_0 = 0$. Also, let

$$F(\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1}, \alpha, \beta) := \mathbb{E}_{(X,Y)\sim\eta(X|Y)\nu(Y)} l(f_{(\alpha,\beta)}(X), Y), \tag{39}$$

where $\eta(X|Y = 1) \sim N(\psi_1, \Sigma_1)$, $\eta(X|Y = -1) \sim N(\psi_{-1}, \Sigma_{-1})$.
Note that $F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha_0, \beta_0) = E_{(X,Y)\sim\nu_{(1-\epsilon)\mu}(X|Y)\nu(Y)} l(f_{(\mu,0)}(X), Y)$.
From **Step 1** of the proof, we have shown that, $F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \geq F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha_0, \beta_0)$.
Now, we claim it is enough to show $F(\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1}, \alpha_0, \beta_0) \leq F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha_0, \beta_0)$ for $\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1} \in A(\nu_\mu, \epsilon)$, since then

$$\begin{aligned}
& \max_{\tilde{A}(\nu_\mu,\epsilon)} inf_{\alpha,\beta} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \\
\geq\ & inf_{\alpha,\beta} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \\
\geq\ & F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha_0, \beta_0) \\
\geq\ & \max_{\tilde{A}(\nu_\mu,\epsilon)} F(\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1}, \alpha_0, \beta_0) \\
\geq\ & inf_{\alpha,\beta} \max_{\tilde{A}(\nu_\mu,\epsilon)} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta)
\end{aligned} \tag{40}$$

and we know that

$$inf_{\alpha,\beta} \max_{A(\nu_\mu,\epsilon)} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \geq \max_{A(\nu_\mu,\epsilon)} inf_{\alpha,\beta} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \tag{41}$$

This would imply that

$$\begin{aligned}
inf_{\alpha,\beta} \max_{A(\nu_\mu,\epsilon)} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) &= \max_{A(\nu_\mu,\epsilon)} inf_{\alpha,\beta} F(\psi_1^0, \psi_{-1}^0, \Sigma_1^0, \Sigma_{-1}^0, \alpha, \beta) \\
&= E_{(X,Y)\sim\nu_{(1-\epsilon)\mu}(X|Y)\nu(Y)} l(f_{(\mu,0)}(X), Y)
\end{aligned} \tag{42}$$

and complete the proof.
Now,

$$F(\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1}, \alpha_0, \beta_0) = P\left(Z < \frac{\mu'\psi_{-1}}{\sqrt{\mu'\Sigma_{-1}\mu}}\right) + P\left(Z > \frac{\mu'\psi_1}{\sqrt{\mu'\Sigma_{-1}\mu}}\right). \tag{43}$$

Since the constraints in $A(\nu_\mu, \epsilon)$ are separate and symmetric with respect to ψ_1, Σ_1 and ψ_{-1}, Σ_{-1} respectively, and standard Normal distribution is symmetric. Therefore the max of $F(\psi_1, \psi_{-1}, \Sigma_1, \Sigma_{-1}, \alpha_0, \beta_0)$ over $A(\nu_\mu, \epsilon)$ is attained when $\psi_{-1} = -\psi_1$ and $\Sigma_1 = \Sigma_{-1}$. The remainder follows from **Step 1**.

A.5. Error probabilities in closed form (Definition 4.2)

Standard accuracy probability. The standard accuracy probability is given as

$$\mathbb{P}_{x\sim D}(f_\theta(x) = y) = 1 - \mathbb{P}_{x\sim D}(yx^\top\theta < 0) = 1 - Q\left(\frac{\mu^\top\theta}{\sigma\|\theta\|}\right) \tag{44}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$.

l_∞^ϵ robust accuracy. The l_∞^ϵ robust accuracy is given as:

$$\begin{aligned}\inf_{\|v\|_\infty<\epsilon} \mathbb{P}_{x\sim D}(f_\theta(x)=y) &= 1-\mathbb{P}_{x\sim D}\left(\inf_{\|v\|_\infty<\epsilon}\{y\cdot(x+v)^\top\theta\}<0\right)\\ &= 1-Q\left(\frac{\mu^\top\theta}{\sigma\|\theta\|}-\frac{\epsilon\|\theta\|_1}{\sigma\|\theta\|}\right)\end{aligned} \tag{45}$$

$l_2^{\sigma_s}$ smoothed classifier for certifiable robust accuracy. It is the accuracy probability from a $N(0,\sigma_s^2 I)$ smoothed classifier:

$$\mathbb{P}_{x\sim D,\delta\sim N(0,\sigma_s^2 I)}(f_\theta(x+\delta)=y) \tag{46}$$

The above probability involves two random variables $x(y)\sim D$ and $\delta\sim N(0,\sigma_s^2 I)$, which are independent of each other. This addition is considered as the convolution of two pdfs. Since D is a mixture of Gaussian and δ subjects to a Gaussian, there is a closed form

$$D^* = N(\mu,\sigma^2 I) * N(0,\sigma_s^2 I) = N(\mu,\sigma^2+\sigma_s^2 I) \tag{47}$$

Thus, the certifiable robust accuracy in this setting is equivalent to the standard accuracy probability from a convoluted dataset.

$$\mathbb{P}_{x\sim D^*}(f_\theta(x)=y) = 1-Q\left(\frac{\mu^\top\theta}{(\sigma+\sigma_s)\|\theta\|}\right) \tag{48}$$

A.6. Provable improvement from data augmentation (Theorem 4.3)

Proof. We divide the proof into several steps as follows:

Step 1: The distribution of original data (X,Y) is iid with $X|Y\sim N(y\mu,\sigma^2 I)$. Thus, from Eq.(10) and (11) in (Carmon et al., 2019), we have

$$\mathbb{PE}_a^{p,\epsilon}(f_\theta) = Q\left(\frac{\mu^T\theta}{\sigma\|\theta\|}-\frac{\epsilon\sqrt{d}}{\sigma}\right),\ \text{where } Q(x)=\frac{1}{2\pi}\int_x^\infty e^{-z^2/2}\mathrm{d}z. \tag{49}$$

Let $\alpha(n_0,n_1)$ be used to denote $n_1/(n_0+n_1)$. Additionally, let $\hat\theta_{n_1} = \dfrac{\left(\sum_{i=1}^{n_1}\tilde Y_i\tilde X_i\right)}{n_1}$.

Step 2: Note that $\mathbb{PE}_a^{p,\epsilon}(f_{\hat\theta_{n_0}}) \geq \mathbb{PE}_a^{p,\epsilon}(f_{\hat\theta_{n_0,n_1}})$ holds as long as $\dfrac{\mu^T\hat\theta_{n_0}}{\|\mu\|\|\hat\theta_{n_0}\|} \leq \dfrac{\mu^T\hat\theta_{n_0,n_1}}{\|\mu\|\|\hat\theta_{n_0,n_1}\|}$. This is what we will show.

Step 3: Let $\hat\theta_{n_1}\in\{x:\|x-t\mu\|\leq A_3 := t\log(n_1)/n_1\}$. Also assume:

$$\hat\theta_{n_0}\in\{x: A_1 := (t+n_0)/\sqrt{n_1}\leq\|x-\mu\|A_2 := \log(n_1)\}\cap\{x:\mu^T(x)/\|x\| \tag{50}$$

$$\leq\sqrt{\|\mu\|^2-((t+n_0)/\sqrt{n_1})^2\}\}} \tag{51}$$

We show that in this case $\dfrac{\mu^T\hat\theta_{n_0}}{\|\mu\|\|\hat\theta_{n_0}\|} \leq \dfrac{\mu^T\hat\theta_{n_0,n_1}}{\|\mu\|\|\hat\theta_{n_0,n_1}\|}$. Indeed,

$$\begin{aligned}\frac{\mu^T\hat\theta_{n_0}}{\|\mu\|\|\hat\theta_{n_0}\|} &\leq \cos(\pi/2-\arcsin(A_1/\|\mu\|))\\ &< \cos\left(\pi_2-\arcsin\left(\frac{n_1A_3+n_0A_2}{n_1t\|\mu\|+n_0\|\mu\|}\right)\right)\\ &< \frac{(n_1t\mu+n_0\mu)^T\hat\theta_{n_0,n_1}}{\|n_1t\mu+n_0\mu\|\|\hat\theta_{n_0,n_1}\|} \qquad (52)\\ &= \frac{\mu^T\hat\theta_{n_0,n_1}}{\|\mu\|\|\hat\theta_{n_0,n_1}\|}. \qquad (53)\end{aligned}$$

The first inequality holds by Cauchy-Schwartz, while the second inequality holds by substituting values of A_1, A_2 and A_3 (note that $n_1A_3 + n_0A_2 \leq (n_1t + n_0)A_1$) combined with the facts that $tN_1 + n_0 > \sqrt{N_1}\log(N_1)$ and $f(n) = tn + n_0 - \sqrt{n}\log(n)$ is monotonically increasing in n.

Note that $\hat{\theta}_{n_0,n_1} = \alpha(n_0,n_1)\hat{\theta}_{n_1} + (1-\alpha(n_0,n_1))\hat{\theta}_{n_0}$. Moreover, $\hat{\theta}_{n_1}$ and $\hat{\theta}_{n_0}$ follow $\hat{\theta}_{n_1} \sim N(t\mu, \sigma^2 I_d/n_1)$ and $\hat{\theta}_{n_0} \sim N(\mu, \sigma^2 I_d/n_0)$, independently.

The third inequality therefore holds because $\|y - t\mu\| \leq A_3$ and $\|x-\mu\| \leq A_2$ implies $\|\alpha(n_0,n_1)y + (1-\alpha(n_0,n_1))x - \alpha(n_0,n_1)t\mu + (1-\alpha(n_0,n_1))\mu\| \leq \alpha(n_0,n_1)A_3 + (1-\alpha(n_0,n_1))A_2$.

Therefore, $P\left(\frac{\mu^T\hat{\theta}_{n_0}}{\|\mu\|\|\hat{\theta}_{n_0}\|} \leq \frac{\mu^T\hat{\theta}_{n_0,n_1}}{\|\mu\|\|\hat{\theta}_{n_0,n_1}\|}\right) \geq P(\hat{\theta}_{n_1} \in \{x : \|x - t\mu\| \leq A_3 := t\log(n_1)/n_1\}) \times P(\hat{\theta}_{n_0} \in \{x : A_1 := (t+n_0)/\sqrt{n_1} \leq \|x-\mu\| \leq A_2 := \log(n_1)\} \cap \{x : \mu^T(x)/\|x\| \leq \sqrt{\|\mu\|^2 - ((t+n_0)/\sqrt{n_1})^2}\}\})$.

The result now holds true by simple variable transformations. □

B. Additional intuition & related works

B.1. Related works

Nonparametric gradient flow (Liutkus et al., 2019) provides satisfactory data generation with theoretical guarantees. Further, data synthesis (Hua et al., 2023) on the feature-Gaussian manifold can be realized by the maximum mean discrepancy (MMD) gradient flow (Mroueh & Nguyen, 2021) is proven to be effective for transfer learning tasks.

Recent studies have shown that data augmentation, including potentially unlabelled data, improves adversarial robustness, both empirically and theoretically. A line of study is rigorously explaining how the adversarial robustness is affected by the sample size (Schmidt et al., 2018) where unlabeled or out-of-distribution data are helpful (Carmon et al., 2019; Deng et al., 2021; Dan et al., 2020; Bhagoji et al., 2019). Although conventional data augmentation techniques (such as random flips, rotations, or clips) (Howard, 2013; He & Sun, 2015) have remarkable standard learning performance, is not until recently that researchers started to investigate how data augmentation can improve deep learning robustness (Volpi et al., 2018; Ng et al., 2020; Rebuffi et al., 2021).

A few studies have already explored the usage of optimal transport (OT) ideas within mixup. PuzzelMix (Kim et al., 2020) aligns saliency regions of images with a masked optimal transport. The idea of augmentation with saliency is extended in recent works (Kim et al., 2021b). OT also helps align feature tensors in high-dimensional vector space (Venkataramanan et al., 2022), while this method also requires autoencoder models to capture reasonable representations, they perform barycentric projection (linear map) and implicitly restrict the interpolation manifold. The idea of barycenter is also used in AutoMix (Zhu et al., 2020), which relies on a barycenter generator. On the other hand, OT has been shown to be effective in various areas such as data augmentation for cardiovascular diseases (Zhu et al., 2022b; Qiu et al., 2022a), model personalization (Zhu et al., 2022a), and multimodal learning (Qiu et al., 2022b).

B.2. Intuitions

The recently proposed idea of randomized smoothing (Cohen et al., 2019b) provides a certifiably robust classifier on ℓ_2-perturbations. This notion is desirable as it not only offers a provable guarantee of the robustness of deep neural networks but also can be obtained in a relatively computationally-efficient fashion. To improve the robustness, multiple approaches have been proposed to facilitate the training of base classifiers that could have better certified robustness. In addition to different regularization methods (Zhai et al., 2020; Jeong & Shin, 2020; Li et al., 2018), the pioneering work adopted Gaussian data augmentation (Cohen et al., 2019b) to pursue smoothness, augmenting more complicated smoothing distribution (Li et al., 2021), adversarial samples (Salman et al., 2019), and mixing adversarial samples with original samples (Jeong et al., 2021) are all promising methods.

We can illustrate our intuitions as follows: (1) Instead of the datapoint-specific adversarial perturbations that are aimed to attack one specific sample, the directed augmented data distribution can be considered as universal perturbations (Moosavi-Dezfooli et al., 2019) that cause label change for a set of samples from the perturbed distribution $\mathcal{U}_P$. (2) Such perturbation matches the global manifold structure of the dataset (Greenewald et al., 2021), therefore promoting a smoother decision boundary. (3) It is shown in (Wei et al., 2020) that this augmentation strategy improves the expansion of the neighborhood of class-conditional distributions. This formulation allows us to employ the results from OT theories (Villani, 2009) and Wasserstein Barycenter (Agueh & Carlier, 2011) thus firmly estimating the perturbed distribution $\mathcal{U}_P$. (4) Apart from most data augmentation techniques that improve the generalization by creating samples that are likely to cover the testing distribution. Our argumentation promotes the inter-class behavior (Tokozume et al., 2017) and potentially enlarges the ratio of the between-class distance to the within-class variance (Fisher, 1936), or Fisher's criterion.

B.3. Mixture of distributions

Mixture of distributions in classification While it is challenging to specify a general distribution family for $\mathcal{U}_P$ to enable robust training, we further look into the structure of the joint data distribution $\mathcal{P}^{all}_{x,y}$. Considering a k-class classification problem, it is natural to view the data distribution as a mixture of subpopulation distributions

$$\mathcal{P}^{all}_{x,y} = \sum_{i=1}^{k} w_k P_k(X, Y), \tag{54}$$

where each mixture $P_k(X, Y)$ stands for an individual class (Carmon et al., 2019; Zhai et al., 2019; Dan et al., 2020). In such a case, we can utilize the geometric structure underneath the training data to specify the adversarial distribution and improve the model's robustness, especially when differentiating subpopulation distributions.

B.4. Connection to Mixup

Let us define $x \in \mathcal{X}$ to be input data and $y \in \mathcal{Y}$ to be the output label. Let $(X_0, Y_0) \sim P^0_{x,y}$ and $(X_1, Y_1) \sim P^1_{x,y}$ be two distributions on $\mathcal{X} \times \mathcal{Y}$. We may also allow $Y_1 = 1$ a.s. $P^1_{x,y}$ and $Y_0 = 0$ a.s. $P^0_{x,y}$.For brevity, we use $P^i_{x,y}$ and P^i interchangeably. For the mixup setting (Kim et al., 2020), the goal is to optimize a loss function as follows:

$$\min_f \mathbb{E}_{(X_0,Y_0)\sim P^0_{x,y},(X_1,Y_1)\sim P^1_{x,y}} \mathbb{E}_{\lambda\sim q}(l(f(h_\lambda(X_0, X_1)), e_\lambda(Y_0, Y_1))) \tag{55}$$

where the label mixup function is $e_\lambda(y_0, y_1) = (1 - \lambda)y_0 + \lambda y_1$. Input mixup uses $h_\lambda(x_0, x_1) = (1 - \lambda)x_0 + \lambda x_1$, and q is the distribution of λ considered. Here conventional mixup (Zhang et al., 2018) considers the linear interpolation of independently sampled data. However, we feel this may be too restrictive in nature, since this may lead to the creation of samples that may not contribute in situations when the task is binary classification. In that respect, consider the following minimization problem instead,

$$\min_f \mathbb{E}_{t\sim q}\mathbb{E}_{(X_0,X_1,Y_0,Y_1)\sim\tilde{\pi}} \left[l(f(h_t(X_0, X_1)), e_t(Y_0, Y_1)\right],$$

where $\tilde{\pi}(x_0, x_1, y_0, y_1) = \tilde{\pi}_x((x_0, x_1)|(y_0, y_1)) \times \tilde{\pi}_y((y_0, y_1))$, with $\tilde{\pi}_x((\cdot, *)|(y_0, y_1))$ being the optimal transport coupling between the conditional distributions $P^0(\cdot|y_0)$ and $P^1(\cdot|y_1)$, while $\tilde{\pi}_y((\cdot, \cdot))$ is the optimal transport coupling between the marginal distributions of Y_0 and Y_1. It is easy to see that under the existence of unique Monge maps in-between marginal and conditional distributions of Y_0, Y_1 and $X_0|Y_0, X_1|Y_1$ respectively, this is equivalent to solving the following optimization problem.

$$\min_f \mathbb{E}_{t\sim q}\mathbb{E}_{(x,y)\sim\mu_t^{\{P_0,P_1\}}} \left[l(f(x), y)\right] \tag{56}$$

where Eq.(2) $\mu_t^{\{P_0,P_1\}} = \min_\mu U^{N=2}_{wb}(\mu) := \min_\mu (1 - t)W(\mu, P^0) + tW(\mu, P^1)$ and $W(\cdot, *)$ is the Wasserstein metric.

B.5. A more general problem formulation: multi-marginal optimal transport

We restrict the adversarial distribution family to be the geodesic interpolations between individual subpopulation distributions. Thus eq.(1) becomes

$$\min_f \max_\alpha \mathbb{E}_{x,y\sim U^{wb}(\alpha)} \left[l(f(x), y)\right] \tag{57}$$

where $U^{wb}(\alpha) = \arg\min_U \frac{1}{K} \sum_i \alpha_i W(U, P_i)$ is the Wasserstein barycenter, in other words, the interpolation of subpopulation distributions.

In addition, with a predictive function f, we can consider a dynamic metric, *geodesic loss*, $\mathrm{R}^{\mathrm{geo}}_f$ that measures the change of its performance with the criteria function l while *gradually interpolating* among subpopulation distributions $\{P_i\}_{i=1}^K$. This metric is thus a continuous function of α, where

$$\mathrm{R}^{\mathrm{geo}}_f(\alpha) = \mathbb{E}_{x,y\sim U^{wb}(\alpha)} \left[l(f(x), y)\right] \tag{58}$$

The geodesic loss $\mathrm{R}^{\mathrm{geo}}_f$ provides us a new lens through which we can measure, interpret, and improve a predictive model's robustness from a geometric perspective.

C. Algorithm & Computation

Algorithm 2 Sinkhorn Barycenter

Input: Empirical distributions α_1, α_2, cost matrix $\mathbf{C}$, $\mathbf{K} = exp(-\mathbf{C}/\epsilon)$.
Output: debiased barycenters α_{S_ϵ}
Initialize scalings $(b_1, b_2), d$ to $\mathbf{1}$.
while not converge **do**
  for $k = 1$ to 2 **do**
    $a_k \leftarrow (\alpha_k / \mathbf{K} n_k)$
  end for
  $\alpha \leftarrow d \odot \sum_{k=1}^{K} (\mathbf{K}^\top a_k)^{w_k}$
  for $k = 1$ to 2 **do**
    $b_k \leftarrow \left(\alpha / \mathbf{K}^\top \alpha_k\right)$
  end for
  $d \leftarrow \sqrt{d \odot (\alpha \mathbf{K} d)}$
end while

In practice, we only observe discrete training samples that represent an empirical distribution of P_i and P_j. Consider $\mathbf{X}_i = \{\mathbf{x}_l^i\}_{l=1}^{n_i}$ and $\mathbf{X}_j = \{\mathbf{x}_l^j\}_{l=1}^{n_j}$ are two set of features from class i and j respectively. The empirical distributions are written as $\hat{P}_i = \sum_{l=1}^{n_i} p_l^i \delta_{x_l^i}$ and $\hat{P}_j = \sum_{l=1}^{n_j} p_l^j \delta_{x_l^j}$ where δ_x is the Dirac function at location $x \in \Omega$, p_l^i and p_l^j are probability mass associated to the sample. Then the Wasserstein distance, between empirical measures $\hat{P}_i$ and $\hat{P}_j$ becomes

$$W(\hat{P}_i, \hat{P}_j) = \inf_{\pi \in \hat{\Pi}} \sum_{l=1,k=1}^{n_i,n_j} c(\mathbf{x}_l^i, \mathbf{x}_k^j) \pi_{l,k} + H(\pi), \tag{59}$$

where $\hat{\Pi} := \{\pi \in (\mathbb{R}^+)^{n_i \times n_j} | \pi \mathbf{1}_{n_j} = \mathbf{1}_{n_i}/n_i, \pi^\top \mathbf{1}_{n_i} = \mathbf{1}_{n_j}/n_j\}$ with $\mathbf{1}_n$ a length n vector of ones, $H(\cdot)$ is the negative entropy regularizer for us to utilize the Sinkhorn algorithm. $c(x, y)$ is the ground cost function that specifies the actual cost to transport the mass, or probability measure, from position x to y. Most studies merely use l_2 norm as the ground metric as there are a lot of desirable properties, such as the linear barycentric projection (Villani, 2009) used in other OT-based mixup methods (Venkataramanan et al., 2022).

Computation concerns: batch OT and entropic OT Discrete optimal transport involves a linear program that has an $O(n^3)$ complexity. Hence, the potential computation issues can not be ignored.
First of all, we adopted the celebrated entropic optimal transport (Cuturi, 2013) and used the Sinkhorn algorithm to solve for OT objectives and Barycenters (Janati et al., 2020) (algorithm 2). The Sinkhorn algorithm has a $O(n \log n)$ complexity, thus it can ease the computation burden. In addition, the pairwise Wasserstein distance can be precomputed and stored. Last but not least, we follow the concept of minibatch optimal transport (Fatras et al., 2021b) where we sample a batch of samples from each condition during the data augmentation procedure. Whereas minibatch OT could lead to non-optimal couplings, our experimental results have demonstrated that our data augmentation is still satisfactory.

Data augmentation In our work, we focus on studying the advantages and limitations brought by certain data augmentation algorithm $\mathcal{A} : \mathcal{X} \mapsto \mathcal{X}$ such that $\mathcal{A}_\# P^0_{X,Y} = P^1_{X,Y}$, where $A_\# \mu = \nu$ denotes that ν is the pushfoward measure of μ by A. Under this notation, a data augmentation algorithm is a transport map that transforms the original data distribution $P^0_{X,Y}$ towards the augmentation distribution $P^1_{X,Y}$. In practice, the estimation of such a transport map is challenging. Nevertheless, we can assume the access to the distribution $P^1_{X,Y}$ and augmented data samples $\{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{n_1} \sim P^1_{X,Y}$. Consider a supervised learning algorithm $T(\cdot, \cdot) : (\mathcal{X} \times \mathcal{X}) \mapsto \Theta$ that maps a dataset to a model parameter θ. The standard training process relies on the raw dataset $\hat{\theta}_{n_0} = T(X, Y)$ while data augmentation provides additional dataset and $\tilde{\theta}_{n_0+n_1} = T([X, Y], [\tilde{X}, \tilde{Y}])$.

D. Additional Experimental Results

D.1. Experiments on MNIST dataseet

In Table 5, we show the comparison results on the MNIST dataset with certified accuracy at various radii, and the comparison results on the CIFAR-10 dataset are shown in Table 4. We set our results **bold-faced** whenever the value improves the Gaussian baseline and underlines whenever the value improves the best among the considered baselines.
As shown in Table 5, our method can significantly improve the certified test accuracy compared with Gaussian (Cohen et al., 2019b) on the MNIST dataset, and also outperforms existing methods, i.e., SmoothAdv (Salman et al., 2019), MACER

Table 5: Certified accuracy on MNIST dataset.

σ	Models (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75
0.25	Gaussian	99.2	98.5	96.7	93.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Stability training	99.3	98.6	97.1	93.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv	99.4	99.0	98.2	96.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER	99.3	98.7	97.5	94.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Consistency	99.3	98.7	98.2	95.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothMix	99.2	98.8	98.0	96.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ours	99.0	98.1	**97.3**	**95.8**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ours + SmoothAdv	98.2	97.1	96.3	**94.7**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ours + SmoothMix	98.3	97.7	**97.0**	**96.1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	Gaussian	99.2	98.3	96.8	94.3	89.7	81.9	67.3	43.6	0.0	0.0	0.0	0.0
	Stability training	99.2	98.5	97.1	94.8	90.7	83.2	69.2	45.4	0.0	0.0	0.0	0.0
	SmoothAdv	99.0	98.3	97.3	95.8	93.2	88.5	81.1	67.5	0.0	0.0	0.0	0.0
	MACER	98.5	97.5	96.2	93.7	90.0	83.7	72.2	54.0	0.0	0.0	0.0	0.0
	Consistency	99.2	98.6	97.6	95.9	93.0	87.8	78.5	60.5	0.0	0.0	0.0	0.0
	SmoothMix	98.7	98.0	97.0	95.3	92.7	88.5	81.8	70.0	0.0	0.0	0.0	0.0
	ours	98.1	97.3	96.2	**94.8**	**92.2**	**87.8**	**79.5**	**67.7**	0.0	0.0	0.0	0.0
	ours + SmoothAdv	88.8	86.7	84.4	80.6	77.6	73.9	**70.3**	**64.0**	0.0	0.0	0.0	0.0
	ours + SmoothMix	97.7	97.0	95.4	93.6	89.1	**84.9**	**78.0**	**67.7**	0.0	0.0	0.0	0.0
1.00	Gaussian	96.3	94.4	91.4	86.8	79.8	70.9	59.4	46.2	32.5	19.7	10.9	5.8
	Stability training	96.5	94.6	91.6	87.2	80.7	71.7	60.5	47.0	33.4	20.6	11.2	5.9
	SmoothAdv	95.8	93.9	90.6	86.5	80.8	73.7	64.6	53.9	43.3	32.8	22.2	12.1
	MACER	91.6	88.1	83.5	77.7	71.1	63.7	55.7	46.8	38.4	29.2	20.0	11.5
	Consistency	95.0	93.0	89.7	85.4	79.7	72.7	63.6	53.0	41.7	30.8	20.3	10.7
	SmoothMix	93.5	91.3	87.9	83.2	77.9	71.1	62.5	53.6	44.9	36.5	28.8	21.3
	ours	91.7	88.7	85.4	81.1	75.4	68.0	**61.4**	**52.3**	**45.0**	**37.8**	**30.7**	**23.2**
	ours+ SmoothAdv	86.2	82.4	78.9	73.9	67.9	62.4	56.8	**49.4**	**43.7**	**38.4**	**33.0**	**27.8**
	ours + SmoothMix	92.5	90.2	86.5	83.0	77.3	70.6	**62.6**	**53.4**	**45.9**	**37.8**	**30.7**	**22.5**

(Zhai et al., 2020), Consistency (Jeong & Shin, 2020), and SmoothMix (Jeong et al., 2021). The important characteristic of our method is the robustness under larger noise levels. Our method achieved the highest certified test accuracy among all the noise levels when the radii are large, i.e., radii 0.50-0.75 under noise level $\sigma = 0.25$, radii 0.75-1.75 under noise level $\sigma = 0.50$, and radii 1.50-2.75 under noise level $\sigma = 1.00$, which clearly demonstrated the effectiveness of our data augmentation method, as the robustness improvement under large noise level is more critical (Cohen et al., 2019a).
We also combined our method with SmoothAdv and SmoothMix to evaluate whether our data augmentation method can provide additive improvements. As shown in Table 5, we find that by combing our data augmentation mechanism, the performance of previous SOTA methods can be even better, which demonstrates the effectiveness of our method, and can be easily used as an add-on mechanism for many other algorithms to improve learning robustness.

D.2. More ablation studies

Comparison on training batch size We show the performance comparison for different training batch sizes on the MNIST dataset.

Table 6: Comparison of different training batch with VAE on the MNIST dataset of certified accuracy at various radii (noise level $\sigma = 0.5$).

σ	Training batch (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.5	64	97.5	96.5	95.2	93.5	90.4	84.5	76.5	64.4	0.0	0.0	0.0
	128	97.7	96.9	96.0	94.5	91.5	87.0	78.7	65.8	0.0	0.0	0.0
	256	97.9	96.8	95.9	94.3	91.6	86.5	79.0	66.3	0.0	0.0	0.0
	512	97.6	97.2	96.1	94.5	91.8	87.0	78.7	67.0	0.0	0.0	0.0
	1024	97.9	97.0	96.1	94.1	91.7	86.7	78.4	66.8	0.0	0.0	0.0
	2048	97.7	97.2	96.1	94.2	91.8	86.4	78.2	67.0	0.0	0.0	0.0
	3072	97.7	96.8	96.0	94.1	91.5	86.2	78.1	66.6	0.0	0.0	0.0
	4096	97.4	96.7	95.4	93.5	91.0	86.0	78.1	66.0	0.0	0.0	0.0

Comparison on augmentation batch size b_n We show the performance comparison for different augmentation batch sizes on the MNIST dataset.

Table 7: Comparison of different augmentation batch size b_n with VAE on the MNIST dataset of certified accuracy at various radii (noise level $\sigma = 0.5$).

σ	augmentation batch b_n (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.5	4	97.9	96.8	95.9	94.4	91.9	87.0	78.7	67.5	0.0	0.0	0.0
	8	97.8	97.2	95.9	94.6	91.6	87.4	79.2	66.7	0.0	0.0	0.0
	16	97.9	97.1	96.1	94.2	91.8	87.8	78.7	67.0	0.0	0.0	0.0
	32	98.0	96.9	95.9	94.8	91.7	87.0	79.2	67.1	0.0	0.0	0.0
	64	97.8	96.9	95.8	94.2	92.1	87.5	78.6	67.5	0.0	0.0	0.0
	128	97.9	97.2	95.9	94.7	91.8	87.5	78.6	66.3	0.0	0.0	0.0
	256	97.7	97.1	96.1	94.4	91.9	88.0	78.8	67.6	0.0	0.0	0.0
	512	97.8	97.2	96.2	94.6	91.8	87.1	79.1	66.8	0.0	0.0	0.0
	1024	98.0	96.9	96.2	94.5	91.1	86.7	78.4	67.3	0.0	0.0	0.0

Influence of magnificent coefficient m_a The comparison results of different magnificent coefficient m_a with corresponding certified accuracy at various radii are shown in Table 8.

Table 8: Comparison of different magnificent coefficients m_a on the MNIST dataset of certified accuracy at various radii. (noise level $\sigma = 0.5$)

σ	magnificent coefficient m_a (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75
0.5	1	97.8	97.2	95.9	94.3	92.2	87.6	78.7	67.5	0.0	0.0	0.0	0.0
	2	98.1	97.3	96.2	94.8	92.2	87.8	78.7	67.7	0.0	0.0	0.0	0.0
	3	98.2	97.3	96.0	94.7	92.0	87.4	78.6	66.6	0.0	0.0	0.0	0.0
	4	97.9	97.1	95.8	94.6	92.2	87.8	78.5	66.8	0.0	0.0	0.0	0.0
	5	98.1	97.1	96.1	94.5	92.3	87.4	79.2	66.4	0.0	0.0	0.0	0.0
	6	97.8	97.0	96.1	94.5	92.6	87.5	79.2	67.0	0.0	0.0	0.0	0.0
	7	98.1	96.7	96.0	94.5	92.0	87.7	78.9	66.4	0.0	0.0	0.0	0.0

Influence of ot_lbd The comparison results of different ot_lbd with corresponding certified accuracy at various radii are shown in Table 9.

Table 9: Comparison of different ot_lbd on the MNIST dataset of certified accuracy at various radii. (noise level $\sigma = 0.5$)

σ	ot_lbd (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75
0.5	1e-1	97.3	96.1	94.6	92.5	88.3	82.4	75.1	63.0	0.0	0.0	0.0	0.0
	1e-2	96.7	96.1	94.6	92.3	88.6	82.4	74.6	64.9	0.0	0.0	0.0	0.0
	1e-3	96.8	96.2	94.4	91.6	88.6	82.5	75.1	63.4	0.0	0.0	0.0	0.0
	1e-4	97.2	96.3	94.3	91.3	87.7	82.6	75.4	64.5	0.0	0.0	0.0	0.0

Influence of ot_t The comparison results of different ot_t with corresponding certified accuracy at various radii are shown in Table 10.

Table 10: Comparison of different ot_t on the MNIST dataset of certified accuracy at various radii. (noise level $\sigma = 0.5$)

σ	ot_t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.5	0.1	97.8	97.2	96.1	94.8	92.1	87.6	78.5	66.5	0.0	0.0	0.0
	0.2	97.7	97.2	95.9	94.3	92.2	87.5	78.5	66.6	0.0	0.0	0.0
	0.3	97.9	97.0	96.2	94.5	91.4	86.5	78.4	67.0	0.0	0.0	0.0
	0.4	97.9	97.1	96.1	93.9	91.3	87.3	78.6	67.2	0.0	0.0	0.0
	0.5	98.0	97.3	95.9	93.8	91.0	86.7	77.7	66.7	0.0	0.0	0.0

D.3. More experimental results

Influence of VAE training The training protocol for VAE can also affect performance. We adopted different VAE training mechanisms and computed the corresponding certified accuracy at various radii, which is shown in Table 11.

Table 11: Comparison of different VAE training methods on the CIFAR-10 dataset of certified accuracy at various radii.

σ	VAE training (CIFAR-10)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75
0.25	Without crop and flip	67.80	60.70	51.50	41.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	With crop and flip	67.20	60.30	53.40	44.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.5	Without crop and flip	50.70	46.10	41.60	36.50	32.90	28.20	22.40	18.20	0.00	0.00	0.00	0.00
	With crop and flip	47.50	44.10	39.70	35.50	30.70	26.50	22.40	18.60	0.00	0.00	0.00	0.00
1.0	Without crop and flip	34.90	31.90	28.20	25.20	23.10	21.30	19.20	17.70	15.60	13.60	11.10	9.70
	With crop and flip	36.40	33.00	29.90	26.90	24.30	21.90	20.00	17.60	14.80	12.90	10.40	8.70

Influence of the hidden size z in VAE More ablation study results on the hidden size z in VAE on the CIFAR-10 dataset are shown in Table 12.

Table 12: Comparison of different hidden sizes z in VAE on the CIFAR-10 dataset of certified accuracy at various radii.

σ	z in VAE (CIFAR10)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75
0.25	30	67.90	59.40	48.80	41.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	64	67.80	60.70	51.50	41.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	90	67.90	59.40	48.90	41.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	128	69.70	63.40	56.00	47.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.5	30	51.30	47.00	41.90	38.70	33.80	28.70	23.90	19.20	0.00	0.00	0.00	0.00
	64	50.70	46.10	41.60	36.50	32.90	28.20	22.40	18.20	0.00	0.00	0.00	0.00
	90	51.30	47.00	41.90	38.70	34.00	28.60	23.80	18.80	0.00	0.00	0.00	0.00
	128	55.40	50.70	45.60	40.60	35.20	30.50	25.70	19.50	0.00	0.00	0.00	0.00
1.0	30	33.80	31.00	28.00	24.60	22.00	19.40	17.30	14.90	13.40	10.40	9.40	8.40
	64	34.90	31.90	28.20	25.20	23.10	21.30	19.20	17.70	15.60	13.60	11.10	9.70
	90	40.90	36.90	34.50	31.50	28.10	24.40	22.50	20.40	16.50	14.00	12.30	9.70
	128	39.00	35.50	31.90	28.20	24.40	21.80	18.50	16.10	13.80	11.20	9.20	7.70

Influence of t on MNIST dataset More ablation study results of different ot_t with PCA and VAE on the MNIST dataset are shown in the following tables.

Table 13: Comparison of different t with PCA on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 0.25$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.25	0.001	99.40	98.50	97.20	95.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.005	99.30	98.30	96.90	94.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	99.40	98.40	97.20	95.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.15	99.50	98.20	97.10	94.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.20	99.20	98.70	97.00	95.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	99.50	98.40	97.00	94.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.30	99.30	98.10	97.30	95.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.35	99.40	98.60	96.90	94.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.40	99.60	98.20	96.90	95.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.45	99.40	98.30	97.00	94.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.50	99.40	98.40	97.20	95.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 14: Comparison of different t with VAE on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 0.25$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.25	0.001	99.30	98.60	96.40	93.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.005	99.30	98.20	97.20	93.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	98.90	98.20	97.00	93.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.15	99.10	98.50	97.40	94.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.20	98.80	98.50	97.30	94.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	99.30	98.60	96.80	94.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.30	99.10	98.00	96.80	93.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.35	99.00	98.40	97.00	94.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.40	99.20	98.00	96.90	94.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.45	99.10	98.20	96.60	94.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.50	99.30	98.40	97.00	94.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 15: Comparison of different t with PCA on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 0.5$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.5	0.001	99.00	98.10	96.90	95.40	91.40	85.50	75.10	56.20	0.00	0.00	0.00
	0.005	98.90	97.90	97.10	95.50	92.20	85.70	75.20	57.50	0.00	0.00	0.00
	0.10	99.00	98.30	97.20	95.30	92.10	85.60	75.70	56.50	0.00	0.00	0.00
	0.15	99.10	98.20	97.10	95.50	91.50	85.30	74.80	55.70	0.00	0.00	0.00
	0.20	99.00	98.40	96.70	95.20	91.90	86.00	75.30	58.20	0.00	0.00	0.00
	0.25	99.10	98.10	97.10	95.20	92.10	85.80	75.70	58.00	0.00	0.00	0.00
	0.30	99.10	98.30	97.40	95.30	91.30	86.00	74.60	57.50	0.00	0.00	0.00
	0.35	99.10	97.80	97.10	95.20	92.10	86.30	75.50	57.50	0.00	0.00	0.00
	0.40	99.10	98.10	97.00	95.10	91.60	85.90	75.70	55.90	0.00	0.00	0.00
	0.45	99.00	98.30	97.30	95.60	92.50	85.90	75.00	57.30	0.00	0.00	0.00
	0.50	99.00	98.30	96.70	95.40	91.90	85.60	75.00	57.00	0.00	0.00	0.00

Table 16: Comparison of different t with VAE on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 0.5$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.5	0.001	99.00	98.100	96.90	95.40	91.40	85.50	75.10	56.20	0.00	0.00	0.00
	0.005	98.90	97.90	97.10	95.50	92.20	85.70	75.20	57.50	0.00	0.00	0.00
	0.10	99.00	98.30	97.20	95.30	92.10	85.60	75.70	56.50	0.00	0.00	0.00
	0.15	99.10	98.20	97.10	95.50	91.50	85.30	74.80	55.70	0.00	0.00	0.00
	0.20	99.00	98.40	96.70	95.20	91.90	86.00	75.30	58.20	0.00	0.00	0.00
	0.25	99.100	98.10	97.10	95.20	92.10	85.80	75.70	58.00	0.00	0.00	0.00
	0.30	99.100	98.30	97.40	95.30	91.30	86.00	74.60	57.50	0.00	0.00	0.00
	0.35	99.10	97.80	97.10	95.20	92.10	86.30	75.50	57.50	0.00	0.00	0.00
	0.40	99.10	98.10	97.00	95.10	91.60	85.90	75.70	55.90	0.00	0.00	0.00
	0.45	99.00	98.30	97.30	95.60	92.50	85.90	75.00	57.30	0.00	0.00	0.00
	0.50	99.00	98.30	96.70	95.40	91.90	85.60	75.00	57.00	0.00	0.00	0.00

Table 17: Comparison of different t with PCA on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 1.0$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
1.0	0.001	95.90	93.90	90.20	86.40	80.40	72.70	61.70	50.80	38.80	27.50	16.40
	0.005	96.00	94.00	90.20	86.40	80.40	72.80	61.40	50.90	38.40	28.20	16.50
	0.10	95.70	93.80	90.00	85.90	80.10	72.50	61.60	50.90	39.20	28.20	16.90
	0.15	95.60	93.90	89.80	86.20	80.30	72.80	62.10	50.80	38.10	27.20	15.90
	0.20	96.20	94.00	90.40	86.30	80.30	72.40	61.50	50.40	38.40	27.40	16.30
	0.25	96.10	93.90	90.50	86.10	80.20	72.70	61.40	50.80	38.70	27.90	16.00
	0.30	96.10	94.10	90.00	86.70	80.00	72.40	61.40	50.50	38.10	27.20	16.10
	0.35	95.90	93.80	90.20	86.40	80.30	72.90	61.40	50.70	37.80	27.600	15.800
	0.40	96.10	93.90	90.20	86.60	79.90	72.50	61.50	50.30	38.10	28.00	16.30
	0.45	95.80	93.90	90.10	86.60	80.40	72.30	61.30	50.80	37.80	27.40	16.30
	0.50	96.10	93.60	90.00	86.00	79.80	73.20	61.70	50.90	38.30	27.30	16.10

Table 18: Comparison of different t with VAE on the MNIST dataset of certified accuracy at various radii (without mixing label, noise level $\sigma = 1.0$).

σ	t (MNIST)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
1.0	0.001	96.00	94.20	90.40	86.70	79.70	73.10	61.80	49.90	35.90	24.20	13.30
	0.005	95.70	93.90	90.50	86.20	80.20	72.40	61.80	49.80	35.80	24.90	13.50
	0.10	95.60	93.90	90.00	86.80	80.30	72.60	61.20	50.30	36.00	24.80	13.70
	0.15	95.80	94.00	90.30	86.70	80.30	72.30	61.90	49.90	6.60	25.40	13.70
	0.20	95.50	94.10	89.80	86.20	79.70	72.30	62.00	50.10	37.10	26.100	14.100
	0.25	95.80	93.80	90.10	86.50	80.30	72.70	62.50	50.50	37.10	25.90	14.40
	0.30	95.80	93.70	90.50	86.00	80.30	72.70	62.10	50.00	37.90	26.60	14.50
	0.35	95.50	93.50	89.80	86.10	80.40	72.40	61.70	50.50	37.30	26.50	14.60
	0.40	95.80	93.60	89.80	85.60	79.70	72.30	62.10	51.00	37.80	26.80	14.80
	0.45	95.60	93.20	89.70	86.00	79.80	72.20	62.30	50.10	38.40	28.20	15.70
	0.50	95.50	93.30	89.60	85.80	79.70	72.20	61.80	51.20	38.40	27.80	16.90