Certified Adversarial Robustness Under the Bounded Support Set

Yiwen Kou [1] **Qinyuan Zheng** [2] **Yisen Wang** [2 3]

Abstract

Deep neural networks (DNNs) have revealed severe vulnerability to adversarial perturbations, beside empirical adversarial training for robustness, the design of provably robust classifiers attracts more and more attention. Randomized smoothing methods provide the certified robustness with agnostic architecture, which is further extended to a provable robustness framework using f-divergence. While these methods cannot be applied to smoothing measures with bounded support set such as uniform probability measure due to the use of likelihood ratio in their certification methods. In this paper, we generalize the f-divergence-based framework to a Wasserstein-distance-based and total-variation-distance-based framework that is first able to analyze robustness properties of bounded support set smoothing measures both theoretically and experimentally. By applying our methodology to uniform probability measures with support set $l_p(p = 1, 2, \infty$ and general) ball, we prove negative certified robustness properties with respect to $l_q(q = 1, 2, \infty)$ perturbations and present experimental results on CIFAR-10 dataset with ResNet to validate our theory. And it is also worth mentioning that our certification procedure only costs constant computation time.

1. Introduction

Vulnerability to adversarial samples is a major obstacle that various classifiers obtained by machine learning algorithms, especially deep neural networks (DNNs), need to overcome (Szegedy et al., 2013; Nguyen et al., 2015). For instance, in computer vision applications, deliberately adding some subtle perturbation δ that humans cannot perceive to the input image x will cause DNNs to give a wrong classification output with high probability. Many empirical adversarial defenses have been proposed, among which adversarial training (Madry et al., 2018) is the most effective one (Athalye et al., 2018). However, it still faces stronger or adaptive attacks to decrease its effectiveness to a certain degree (Croce & Hein, 2020). This motivates research on certified robustness: algorithms that are provably robust to the worst-case attacks.

Some works propose algorithms to learn DNNs that are provably robust against norm-bounded adversarial perturbations by using some convex relaxation methods (Wong et al., 2018; Wong & Kolter, 2018; Weng et al., 2018; Mirman et al., 2018; Wang et al., 2018; Singh et al., 2018; Ryou et al., 2020; Zhang et al., 2018; Balunovic & Vechev, 2019; Xiao et al., 2018; Raghunathan et al., 2018; Dvijotham et al., 2020b; Mirman et al., 2019; Dvijotham et al., 2018a;b). However, these approaches are usually computationally expensive and require extensive knowledge of classifier architecture.

Besides, randomized smoothing has received significant attention in recent years for verifying the robustness of classifiers (Liu et al., 2018; Cao & Gong, 2017; Lecuyer et al., 2019; Cohen et al., 2019; Salman et al., 2019; Zhai et al., 2020). Based on this method, several papers have studied which smoothing strategies perform better for specific l_p perturbations. Cohen et al. (2019) conclude that randomized smoothing can be well understood for the l_2 case by using Gaussian probability measure for smoothing. And several special cases of the conjecture have been proven for $p < 2$: Li et al. (2018) show that l_1 robustness can be achieved with the Laplacian distribution, and Lee et al. (2019) show that l_0 robustness can be achieved with a discrete distribution.

Other papers start from the opposite perspective and focus on studying under specific assumptions which perturbation is provably difficult to handle and which smoothing methods are ineffective for particular disturbance. As for the existence of a noise distribution that works for the case of $p > 2$, Blum et al. (2020); Kumar et al. (2020) show hardness results for random smoothing to achieve l_p certified robustness. And Yang et al. (2020) proved that the "optimal" smoothing distributions for any "nice" norms have level sets given by the norm's Wulff Crystal.

[1] Yuanpei College, Peking University [2] Key Lab. of Machine Perception (MoE), School of Artificial Intelligence, Peking University. [3] Institute for Artificial Intelligence, Peking University.. Correspondence to: Yisen Wang <yisen.wang@pku.edu.cn>.

Proceedings of the 39^{th} *International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Notably, based on randomized smoothing strategy, Dvijotham et al. (2020a) introduce a robustness framework by utilizing convex relaxation technique for f-divergence. Notice that the definition of f-divergence is related to the likelihood ratio $r(X) = \frac{\nu(X)}{\rho(X)}$, and $r(X)$ is well-defined only when the support set of ρ contains the support set of ν. Thus, when the support set of reference measure ρ is bounded, and ν takes even a small translation of ρ, the support set of ν will cross over the boundary of support set of ρ. Therefore, the certification of smoothing probability measures with bounded support sets is beyond the discussion of Dvijotham et al. (2020a)'s framework. In this paper, by using Wasserstein distance as well as total variation distance, we provide a robustness certification framework that is first able to analyze robustness properties of bounded support set smoothing measures both theoretically and experimentally. Our contributions are summarized as follows:

- We provide robustness certification results for smoothed classifiers following the setting of Dvijotham et al. (2020a)'s f-divergence-based framework. Here we generalize their framework by taking into consideration a relaxation from an intersection of Wasserstein-distance-based and total-variation-distance-based balls and derive a theoretical framework which is first able to provide certification formulas for bounded support set smoothing measures and hence able to verify robustness properties of bounded support set smoothing measures by experiments.

- By applying our methodology to uniform probability measures with l_1 ball support set, we show theoretically its bad performance for l_2 and l_∞ robustness certification task in theorem 4.3; By applying our methodology to uniform probability measures with l_2 ball support set, we obtain certified robustness properties with respect to l_q-perturbations in theorem 4.4 and theoretically predict its negative performance for $l_q (q > 2)$ adversary; By applying our methodology to uniform probability measures with l_∞ ball support set, we obtain certified robustness properties with respect to l_1, l_2, l_∞-perturbations in theorem 4.5 and show theoretically its poor performance for l_2 and l_∞ adversaries. Furthermore, we analyze the cases when smoothing measure is taken as uniform probability measure with more general support set l_p ball and show the unavoidable curse of dimension for the usage of such smoothing measures.

- We present experimental results on CIFAR-10 dataset with ResNet model to validate part of our theory about uniform smoothing measures with l_2 ball and l_∞ ball support set on l_2 adversary and use Gaussian smoothing measure as contrast. It is worth mentioning that our certification procedure only costs constant computation time.

2. Related Works

Certified Robustness for Conventional Networks. Many recent works focus on certifying the robustness of learned neural networks under any attack. Some works bounded the certified radius of conventional neural networks layer by layer by utilizing some convex relaxation methods, including linear relaxation (Wong et al., 2018; Wong & Kolter, 2018; Weng et al., 2018; Mirman et al., 2018; Wang et al., 2018; Singh et al., 2018; Ryou et al., 2020; Zhang et al., 2018; Balunovic & Vechev, 2019; Xiao et al., 2018), semidefinite relaxation (Raghunathan et al., 2018; Dvijotham et al., 2020b) and interval bound relaxation (IBP) (Mirman et al., 2019; Dvijotham et al., 2018a;b). However, such approaches encountered several drawbacks, such as computationally expensive, incapability to deal with deep and large models and loose bounds which results in unstable training.

Randomized Smoothing. Randomized smoothing was first proposed as a heuristic defense without any guarantees (Liu et al., 2018; Cao & Gong, 2017). It followed that l_1 and l_2 robustness guarantees for smoothing with Gaussian and Laplace noise, respectively, was proposed by Lecuyer et al. (2019) from a differential privacy perspective and a stronger l_2 robustness guarantees for Gaussian noise was proposed by Li et al. (2018) based on information theory. Cohen et al. (2019); Salman et al. (2019); Zhai et al. (2020) provided l_2 robustness guarantees for Gaussian smoothed classifiers.

Curse of Dimensionality. Blum et al. (2020); Kumar et al. (2020) show hardness results for randomized smoothing to achieve certified robustness for $l_p (p > 2)$ perturbations. Nevertheless, since these works provide hardness results for every possible base classifier including those unusual and even bizarre ones, hardness results given by these papers might be over-tight and attributed to taking into account classifiers that will never appear in real-world applications. From this perspective, the order of difficulty restricted within the common classifiers subset still remains unresolved. Yang et al. (2020) proposed a theoretical framework based on the norm's Wulff Crystal and showed that randomized smoothing cannot achieve nontrivial certified accuracy against perturbations of l_p-norm $\Omega(\min\{1, d^{\frac{1}{p}-\frac{1}{2}}\})$ and performed experiments comparing performance between Gaussian, Laplace, Exponential, PowerLaw and Uniform smoothing distribution for l_1, l_2, l_∞ adversary.

F-divergence-based Framework. Notably, based on randomized smoothing strategy, Dvijotham et al. (2020a) introduce a provable robustness framework using f-divergence as their convex relaxation technique. However, due to the use of likelihood ratio in their certification methods, the framework cannot be applied to smoothing measures with bounded support sets such as uniform probability measures. Another related work is Zhang et al. (2020), which propose a general framework of adversarial certification from a unified functional optimization perspective. In this paper, we introduce a framework that is able to deal with robustness properties of arbitrary smoothing measures, including those with bounded support set, by using Wasserstein distance as well as total variation distance.

3. Problem Setting

Given a binary base classifier $h : \mathbb{R}^d \to \mathcal{Y} = \{\pm 1\}$ and smoothing probability measure μ, the randomly smoothed classifier $h_\mu(x)$ is defined as follows.

Definition 1 (smoothed classifier, smoothing measure)**.** *The smoothed version of a base binary classifier h producing labels in set $\mathcal{Y} = \{\pm 1\}$ is defined as*

$$h_\mu(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{X \sim x+\mu}[h(X) = y],$$

where $\mu \in \mathcal{P}(\mathcal{X})$ is called smoothing measure.

Another way to understand this definition is to say that the smoothed classifier first scores point x as $h_{\mu,y}(x) = \mathbb{P}_{X \sim x+\mu}[h(X) = y]$ for each specific class $y \in \mathcal{Y}$ and then outputs the class y^* with the highest score. We want to study the robustness of the smoothed classifier h_μ against adversarial perturbations of size at most ϵ with respect to a given norm $\|\cdot\|_q$. The question that whether a bounded l_q norm adversarial attack on a fixed input x satisfying $h_\mu(x) = +1$ is successful or not can be formulated as solving the optimization problem below:

$$\min_{\|x'-x\|_q \le \epsilon} \mathbb{P}_{X \sim x'+\mu}[h(X) = +1].$$

The attack is successful if and only if the minimum value is smaller than $\frac{1}{2}$. Since we know little about the information of the black-box classifier h, we follow the approach introduced in Dvijotham et al. (2020a): rather than studying the adversarial attack in the input space $\mathcal{X}$, we study it in the space of probability measures defined on input space $\mathcal{P}(\mathcal{X})$,

$$\begin{aligned}&\min_{\|x'-x\|_q \le \epsilon} \mathbb{P}_{X \sim x'+\mu}[h(X) = +1]\\ =&\min_{\nu \in \mathcal{D}_{x,\epsilon,q}} \mathbb{P}_{X \sim \nu}[h(X) = +1],\end{aligned}$$

where $\mathcal{D}_{x,\epsilon,q} := \{x' + \mu : \|x - x'\|_q \le \epsilon\}$ represents an l_q-norm-based constraint set of radius ϵ for smoothing measure μ centered at a particular sample point x. Then, we follow the full-information robust certification framework established in Dvijotham et al. (2020a) and analyze the generalization of binary classifier h, which they called specification and denote it as $\phi : \mathcal{X} \subseteq \mathbb{R}^d \to \mathcal{Z} \subseteq \mathbb{R}$. Besides, we define reference probability measure ρ as $x + \mu$ and a collection of perturbed probability measures $\mathcal{D}_\rho \subseteq \mathcal{P}(\mathcal{X})$, where $\mathcal{D}_\rho$ means that it is related to ρ. Checking whether a given specification ϕ is robustly certified at ρ with respect to $\mathcal{D}_\rho$ or not is equivalent to estimating the optimal value of following optimization problem is non-negative or not:

$$OPT(\phi, \mathcal{D}_\rho) := \min_{\nu \in \mathcal{D}_\rho} \mathbb{E}_{X \sim \nu}[\phi(X)].$$

And certifying l_p robustness on input x with output of smoothed classifier $h_\mu(x) = +1$ is equivalent to verify whether $OPT(h, x + \mu, \mathcal{D}_{x,\epsilon,q}) \ge 0$ or not.

4. Certification Procedures

Since the set of measures $\mathcal{D}_{x,\epsilon,q}$ constraint in optimization problem $OPT(h, x + \mu, \mathcal{D}_{x,\epsilon,q})$ is intractable to deal with, we consider relaxations of this by using Wasserstein distance as well as total variation distance constraints between ν and $x + \mu$, i.e. $\mathcal{D}_{x,\epsilon,q} \subseteq \{\nu : W_p(x + \mu, \nu) \le \delta\} := \mathcal{D}_{x,\delta,p}$ which represents W_p-distance-based constraint set of radius δ for smoothing measure μ centered at sample point x and $\mathcal{D} \subseteq \{\nu : TV(x + \mu, \nu) \le \xi\} := \mathcal{D}_{x,\xi}$ which represents TV-distance-based constraint set of radius ξ for smoothing measure μ centered at sample point x. Combining the two relaxations, we know $\mathcal{D}_{x,\epsilon,q} \subseteq \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi}$ and therefore

$$OPT(h, x + \mu, \mathcal{D}_{x,\epsilon,q}) \ge OPT(h, x + \mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi}).$$

Thus, for a fixed input x, it suffices to consider the Wasserstein distance and total variation distance relaxed problem and verify whether $OPT(h, x + \mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi}) \ge 0$ or not. The analysis of this problem can be divided into three parts: (1) Compute the Wasserstein distance relaxation measure set $\mathcal{D}_{x,\delta,p}$. And we obtain an δ for general probability measure in section 4.1. (2) Compute the total variation distance relaxation measure set $\mathcal{D}_{x,\xi}$. And we obtain several ξ for Gaussian probability measure and simple examples of uniform probability measure with l_1, l_2, l_∞ and general l_p ball support set in section 4.2 for better understanding of our framework; (3) Compute the Lagrange function as well as dual problem of the relaxed optimization problem $OPT(h, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$ and consequently obtain certification formulas for Gaussian smoothing measure, uniform smoothing measure with l_2, l_∞ ball support sets in table 1. The details are discussed in the following three sections.

4.1. Relaxation Using Wasserstein Distance

In this section, we show the following relaxation from norm-based constraint sets into Wasserstein-distance-based constraint sets for general smoothing measures as well as Gaussian smoothing measure. For simplicity, we denote Wasserstein distance as W distance and denote Wasserstein distance with specified parameter p as W_p distance. The main idea of this subsection can be formulated as follows:

$$\begin{aligned}\mathcal{D}_{x,\epsilon,q} &= \{x' + \mu : \|x - x'\|_q \le \epsilon\} \\ &\subseteq \{\nu : W_p(x + \mu, \nu) \le \delta\} = \mathcal{D}_{x,\delta,p}.\end{aligned}$$

Note the difference of definition between $\mathcal{D}_{x,\epsilon,q}$ and $\mathcal{D}_{x,\delta,p}$ that q is the norm of adversary and p is the parameter of Wasserstein distance, while ϵ is the magnitude of l_q adversary and δ is the radius of Wasserstein relaxation set. We want to find a δ as small as possible, which is related to ϵ, q, p and satisfies the above inclusion relation.

4.1.1. General Probability Measure

Here, we want to find a $\delta_q(\epsilon)$ such that

$$\begin{aligned}\mathcal{D}_{x,\epsilon,q} &= \{x' + \mu : \|x - x'\|_q \le \epsilon\} \\ &\subseteq \{\nu : W_p(x + \mu, \nu) \le \delta_q(\epsilon)\} \\ &= \mathcal{D}_{x,\delta_q(\epsilon),p} \text{ for all } \mu \in \mathcal{P}(\mathcal{X}).\end{aligned}$$

Theorem 4.1. *For all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, norm-based constraint set $\mathcal{D}_{x,\epsilon,q}$ can be relaxed into W-distance-based constraint set $\mathcal{D}_{x,\delta_q(\epsilon),p}$ with radius $\delta_q(\epsilon) = \max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}$.*

And this relaxation radius $\max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}$ works for any Wasserstein distance parameter $p > 0$ as well as any smoothing measure μ.

Note that for $l_q(q \le 2)$ adversarial perturbations, the relaxed radius avoids the influence of dimension d, whereas for $l_q(q > 2)$ adversarial perturbations, as q increases, $\frac{1}{2} - \frac{1}{q}$ increases from 0 to $\frac{1}{2}$ correspondingly. The fact that the radius of W_q-distance constraint set grows with order $\Theta(d^{\frac{1}{2}-\frac{1}{q}})$ provides us with an intuition that it is increasingly harder to bound $\mathcal{D}_{x,\epsilon,q}$ with larger q, therefore, W-distance-relaxation works better for $l_q(q \le 2)$ norm perturbation. And this relaxation radius is tight for W_2 distance and Gaussian smoothing measures which is proved in the appendix E and therefore shows that W_2-distance-relaxation works well for Gaussian smoothing measure.

4.2. Relaxation Using Total Variation Distance

In this section, we show the following relaxation from norm-based constraint sets into total-variation-distance-based constraint sets for Gaussian and uniform smoothing measures. For simplicity, we denote total variation distance as TV distance. The main idea of this subsection can be formulated as follows:

$$\begin{aligned}\mathcal{D}_{x,\epsilon,q} &= \{x' + \mu : \|x - x'\|_q \le \epsilon\} \\ &\subseteq \{\nu : TV(x + \mu, \nu) \le \xi\} = \mathcal{D}_{x,\xi}.\end{aligned}$$

Note the difference of definition between $\mathcal{D}_{x,\epsilon,q}$ and $\mathcal{D}_{x,\xi}$ that ϵ is the norm of l_q adversary and ξ is the radius of total variation distance relaxation set. We want to find a ξ as small as possible, which is related to ϵ, q and satisfies the above inclusion relation.

4.2.1. Gaussian Probability Measure

Here, we want to find a $\xi(\epsilon)$ for Gaussian measure $\mu = \mathcal{N}(0, \sigma^2 I)$ such that

$$\begin{aligned}\mathcal{D}_{x,\epsilon,q} &= \{x' + \mu : \|x - x'\|_q \le \epsilon\} \\ &\subseteq \{\nu : TV(x + \mu, \nu) \le \xi(\epsilon)\} = \mathcal{D}_{x,\xi(\epsilon)}.\end{aligned}$$

The magnitude of $\xi(\epsilon)$ is given by the following theorem.

Theorem 4.2. *For Gaussian probability measure $\mu = \mathcal{N}(0, \sigma^2 I)$ on Euclidean space $\mathbb{R}^d$ and for all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, norm-based constraint set $D_{x,\epsilon,q}$ can be relaxed into TV-distance-based constraint set $\mathcal{D}_{x,\xi(\epsilon)}$ with radius $\xi(\epsilon) = 2G\big(\frac{\max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}}{2\sigma}\big) - 1$ where G is the cumulative distribution function for standard normal distribution $\mathcal{N}(0, 1)$.*

This theorem theoretically shows that TV distance relaxation works effectively for $l_q(q \le 2)$ perturbation due to the irrelevance of the radius to dimension d and increasingly bad for $l_q(q > 2)$ perturbation because of the dependence of the radius to dimension d as order $\Theta(d^{\frac{1}{2}-\frac{1}{q}})$.

4.2.2. Uniform Probability Measure

Here, we want to find a $\xi(\epsilon)$ for uniform measure $\mu = \mathcal{U}(K)$, where K is a specific convex compact set in $\mathbb{R}^d$, and $\mathcal{U}(K)$ is a uniform probability measure supported on $K \subseteq \mathbb{R}^d$ with density function $f_K(x) = \frac{1}{\text{Vol}(K)}\mathcal{I}_{x \in K}$ such that

$$\begin{aligned}\mathcal{D}_{x,\epsilon,q} &= \{x' + \mu : \|x - x'\|_q \le \epsilon\} \\ &\subseteq \{\nu : TV(x + \mu, \nu) \le \xi(\epsilon)\} = \mathcal{D}_{x,\xi(\epsilon)}.\end{aligned}$$

In this paper, we mainly focus on the case when K is l_p-norm ball centered at original point O with radius r. We give following theorems about special cases when $p = 1, 2, \infty$.

Theorem 4.3. *When K is an l_1 norm ball centered at O with radius r in $\mathbb{R}^d$ and for all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, for uniform probability measure $\mathcal{U}(K)$ on Euclidean space $\mathbb{R}^d$, we have*

$$\begin{aligned}&\mathcal{D}_{x,\epsilon,q} \setminus \big\{\nu : TV(x + \mu, \nu) \le 1 - \xi'\big\} \ne \emptyset \\ &\textit{for all } q > 1 \textit{ and arbitrarily small } \xi' > 0,\end{aligned}$$

when $\epsilon \ge 2rd^{\frac{1}{q}-1}$.

Note that $\epsilon \ge \frac{2r}{\sqrt{d}}$ for $q = 2$, which decays with order $\Theta(d^{-\frac{1}{2}})$, and $\epsilon \ge \frac{2r}{d}$ for $q = \infty$, which decays with order $\Theta(d^{-1})$, this theorem theoretically shows that for uniform smoothing measures with l_1 ball support set, total variation distance fails to relax measure set $\mathcal{D}_{x,\epsilon,q}$ effectively when $q = 2, \infty$. And this will consequently lead to bad performance for l_2 and l_∞ robustness certification task, which can be seen from the following section discussing the importance of TV-distance-based relaxation radius.

Theorem 4.4. *When K is an l_2 (Euclidean) ball centered at O with radius r in $\mathbb{R}^d$, for uniform probability measure $\mathcal{U}(K)$ on Euclidean space $\mathbb{R}^d$ and for all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, when $\epsilon > \min\{2r, 2rd^{\frac{1}{q}-\frac{1}{2}}\}$, norm-based constraint set $\mathcal{D}_{x,\epsilon,q}$ failed to be relaxed into TV-distance-based constraint set which can be formulated as*

$$\mathcal{D}_{x,\epsilon,q} \setminus \{\nu : TV(x+\mu, \nu) \leq 1-\xi'\} \neq \emptyset$$

for all $q > 1$ and arbitrarily small $\xi' > 0$.

And when $\epsilon \leq \min\{2r, 2rd^{\frac{1}{q}-\frac{1}{2}}\}$, norm-based constraint set $D_{x,\epsilon,q}$ can be relaxed into valid TV-distance-based constraint set $\mathcal{D}_{x,\xi(\epsilon)}$ with radius

$$\xi(\epsilon) = 1 - \int_0^{\arccos\left(\frac{\max\{\epsilon, \epsilon d^{1/2-1/q}\}}{2r}\right)} \sin^n(t)dt \Big/ \int_0^{\frac{\pi}{2}} \sin^n(t)dt$$

This theorem shows that for uniform smoothing measures with l_2 ball support set, when $q \leq 2$, relaxation radius is independent of dimension d, whereas when $q > 2$ relaxation radius starts to be bound up with dimension d and the impact of d grows as q increases. To put it another way, total variation distance relaxation performs well for uniform smoothing measures with l_2 ball support set when $q \leq 2$ and increasingly poor when $q > 2$.

Theorem 4.5. *When K is an l_∞ cube centered at O with radius r, for uniform probability measure $\mathcal{U}(K)$ on Euclidean space $\mathbb{R}^d$ and for all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, when $\epsilon > 2r$, norm-based constraint set $\mathcal{D}_{x,\epsilon,q}$ failed to be relaxed into TV-distance-based constraint set which can be formulated as*

$$\mathcal{D}_{x,\epsilon,q} \setminus \{\nu : TV(x+\mu, \nu) \leq 1-\xi'\} \neq \emptyset$$

for all $q > 0$ and arbitrarily small $\xi' > 0$.

And when $\epsilon \leq 2r$, norm-based constraint set $D_{x,\epsilon,q}$ can be relaxed into valid TV-distance-based constraint set $\mathcal{D}_{x,\xi(\epsilon)}$.

When $q = 1$, $\xi(\epsilon)$ can be taken as $\frac{\epsilon}{2r}$.

When $q = 2$, $\xi(\epsilon)$ can be taken as $1 - \left(1 - \frac{\epsilon}{2d^{\frac{1}{2}}r}\right)^d = 1 - e^{-\frac{\epsilon}{2r}d^{\frac{1}{2}}}$ when $0 < \epsilon \leq 2t_n r$ where $\sqrt{\frac{n-1}{n}} \leq t_n < 1$ and t_n approaches 1 at an exponential rate.

When $2t_n r < \epsilon < 2r$, $\xi(\epsilon)$ can be taken as $1 - \left(\frac{d-1+\sqrt{d(\frac{\epsilon}{2r})^2-d+1}}{d}\right)^{d-1}\left(\frac{1-\sqrt{d(\frac{\epsilon}{2r})^2-d+1}}{d}\right)$.

When $q = \infty$, $\xi(\epsilon)$ can be taken as $1 - \left(1 - \frac{\epsilon}{2r}\right)^d$.

As for uniform smoothing measures with l_∞ cube support set, this theorem shows that the performance towards l_1 perturbation turns out to be fine since TV distance relaxation radius $\frac{\epsilon}{2r}$ has nothing to do with dimension d and the dimensional curse is avoided. However, in this case, TV distance relaxation shows incapability to cope with l_2 and l_∞ perturbation in some extent due to the rate of increasing radius tending to 1 as $\Theta(e^{d^{\frac{1}{2}}})$ and $\Theta(e^d)$.

After discussing the special cases when K is an l_∞ cube or an l_2 Euclidean ball, we then consider the general case when K is an l_p ball centered at the original point with radius r and give a lower bound for TV distance relaxation radius in the following theorem.

Theorem 4.6. *When K is an l_p ball centered at O with radius r, for uniform probability measure $\mathcal{U}(K)$ on Euclidean space $\mathbb{R}^d$. Assume for all $x \in \mathbb{R}^d, \epsilon > 0, q > 0$, norm-based constraint set $\mathcal{D}_{x,\epsilon,q}$ can be relaxed into TV-distance-based constraint set $\mathcal{D}_{x,\xi(\epsilon)}$, then*

$$\xi(\epsilon) \geq 2\int_0^{\frac{\epsilon d^{\frac{1}{p}}}{4r(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}} \exp\left(\frac{1}{p} - e\left(2x\Gamma\left(1+\frac{1}{p}\right)\right)^p\right)dx,$$

for all perturbation norm parameter $q > 0$ with high probability when d is sufficiently large.

A way to interpret this theorem is that as p increases and K correspondingly translates from l_1 norm cross-polytope into l_∞ norm cube, the dependence of integral upper limit $\frac{\epsilon d^{\frac{1}{p}}}{4r(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}$ on dimension d is gradually reduced, which theoretically shows by taking all kinds of l_q perturbations into consideration, when $p > q$, in average scale $\mathcal{U}\big(B_p(O,r)\big)$ tends to perform better than $\mathcal{U}\big(B_q(O,r)\big)$ where $B_p(O,r)$ denotes an l_p ball in $\mathbb{R}^d$ centered at O with radius r. From another perspective, we consider a ball with a fixed radius r. As the dimension d of the base Euclidean space increases, fixed proportion of mass concentrates within a slab of width $\Theta(d^{-\frac{1}{p}})$. Thus, intuitively, it is increasingly difficult to bound the perturbed measure set $\mathcal{D}_{x,\epsilon,q}$ by using TV distance and certify as dimension d enlarge and therefore the curse of dimension is unavoidable when we use uniform smoothing measure $\mathcal{U}(K)$ with bounded support set.

4.3. Verifying Full-Information Robust Certification

Based on the above analysis, in this section, we are now prepared to compute the Lagrange function and dual problem of the relaxed optimization problem $OPT(\phi, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$. Here we mainly focus on the case when reference measure $\rho = x + \mu$ and perturbed probability measure ν are absolutely continuous w.r.t. Lebesgue measure λ on $\mathbb{R}^d$, i.e., $\rho, \nu \ll \lambda$ and discard uncommon cases when ρ, ν are

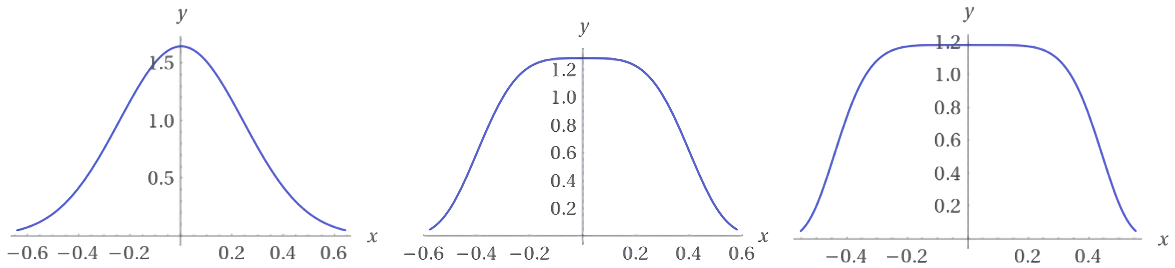

Figure 1. Graph of density function $f(x) = \exp\left(\frac{1}{p} - e(\frac{2}{p}\Gamma(\frac{1}{p}))^p x^p\right)$ when $p = 2, 4, 6$ from left to right

discrete, single or mixed w.r.t. λ. Since $\rho, \nu \ll \lambda$, assume the density function of ρ and ν w.r.t. Lebesgue measure λ are $f(x)$ and $g(x), x \in \mathbb{R}^d$ respectively. Instead of using likelihood ratio $r(x)$, we consider the difference between $g(x)$ and $f(x)$ and define it as $q(x) := g(x) - f(x)$. The objective function $\mathbb{E}_{X\sim\nu}[\phi(X)]$ of optimization problem $OPT(\phi, \rho, \mathcal{D})$ can be rewritten in terms of difference function $q(x)$. And we give the theorems below.

Theorem 4.7 (W_p distance relaxation with $0 < p \leq 1$)**.** *The relaxed optimization problem $OPT(\phi, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$ is equivalent to the convex optimization problem with only one functional variable as below*

$$\begin{aligned} &\inf_{q\in\mathcal{L}^1(\mathcal{X})} \int_{\mathcal{X}} \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)], \\ \textit{s.t.} \quad &\sup_{\|f\|_{L,p}\leq 1} \int f(x)q(x)dx \leq \delta, \int |q(x)|dx \leq 2\xi, \end{aligned} \tag{1}$$

where $\|f\|_{L,p} := \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2^p}$.

Theorem 4.8 (W_p distance relaxation with $p > 1$)**.** *When smoothing measure μ possesses a convex compact support set K and $R := \sup_{y\in K}\|y\|_2, R^* := \|x\|_2 + R + \max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}$, the relaxed optimization problem $OPT(\phi, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$ can be further relaxed into the convex optimization problem with only one functional variable below*

$$\begin{aligned} &\inf_{q\in\mathcal{L}^1(\mathcal{X})} \int_{\mathcal{X}} \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)], \\ \textit{s.t.} \quad &\sup_{\|f\|_L\leq p(2R^*)^{p-1}} \int f(x)q(x)dx \leq \delta^p + (p-1)(2R^*)^{p-1}, \\ &\int |q(x)|dx \leq 2\xi, \end{aligned} \tag{2}$$

where $\|f\|_L := \sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2}$.

Theorem 4.9. *The Lagrange function of optimization problem in* (1) *and* (2) *is*

$$L(\lambda) = \mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\xi - \lambda C, \tag{3}$$

where $\lambda \geq 0$ is the dual variable w.r.t. constraint $\sup_{\|f\|_L\leq 1}\int f(x)q(x)dx \leq \delta$ or constraint $\sup_{\|f\|_L\leq p(2R^*)^{p-1}}\int f(x)q(x)dx \leq \delta^p + (p-1)(2R^*)^{p-1}$ and $C := \delta$ when $0 < p \leq 1$ whereas $C := \delta^p + (p-1)(2R^*)^{p-1}$ when $p > 1$.

Using the duality result, we know the optimal value in (1) can be obtained by computing

$$\max_{\lambda\geq 0} \mathbb{E}_{X\sim x+\mu}[\phi(X)] - \xi - \lambda C = \mathbb{E}_{X\sim x+\mu}[\phi(X)] - \xi, \tag{4}$$

which is only related to the radius ξ of TV distance relaxation set. We can see from this formula the significance of TV distance relaxation radius. By plugging the TV distance relaxation radius given in theorem 4.4, 4.5 and 4.2 in dual optimization problem, we obtain the certification objective in Table 1 and we return certified for l_p norm perturbation with magnitude ϵ if the objective function has non-negative value.

4.4. Relationship with Previous Work

The significance of this paper can be evaluated from two aspects. For the papers about curses of dimensionality (Blum et al., 2020; Kumar et al., 2020), they give no certification procedure and only hardness results which works for every measurable base classifier. However, the classifiers in real-world applications are greatly fewer than measurable ones, which will severely weaken their hardness results. Under such consideration, we provide certification formulas related to base classifier, and in this way, a more practical result can be obtained both theoretically and experimentally. For the papers about the performance of different smoothing measure w.r.t. l_1, l_2, l_∞, l_q adversary (Li et al., 2018; Cohen et al., 2019; Lee et al., 2019; Dvijotham et al., 2020a), the main contribution of our framework is the capability of obtaining certification formula for smoothing measures with bounded support sets, including not only the simplest uniform measures but also the domain-truncated and normalized version of any probability measure μ.

Besides, by applying our methodology to Gaussian probability measure, we miraculously obtain the same certified robustness properties provided in (Dvijotham et al., 2020a) using as Hockey-stick divergence with $\beta = 1$.

Table 1. Certification objectives and prerequisites.

Smoothing Measure	Perturbation	Certification Objective	Prerequisite
$\mathcal{U}(B_2(O,r))$	$l_q(q \leq 2)$	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\Big(1 - \frac{\int_0^{\arccos(\frac{\epsilon}{2r})} \sin^n(t)dt}{\int_0^{\frac{\pi}{2}} \sin^n(t)dt}\Big)$	$\epsilon \leq 2r$
	$l_q(q > 2)$	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\Bigg(1 - \frac{\int_0^{\arccos(\frac{\epsilon d^{1/2-1/q}}{2r})} \sin^n(t)dt}{\int_0^{\frac{\pi}{2}} \sin^n(t)dt}\Bigg)$	$\epsilon \leq 2rd^{\frac{1}{q}-\frac{1}{2}}$
$\mathcal{U}(B_\infty(O,r))$	l_1	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - \frac{\epsilon}{r}$	$\epsilon \leq 2r$
	l_2	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\Big(1 - \big(1 - \frac{\epsilon}{2d^{\frac{1}{2}}r}\big)^d\Big)$	$\epsilon \leq 2t_n r$
	l_∞	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\Big(1 - \big(1 - \frac{\epsilon}{2r}\big)^d\Big)$	$\epsilon \leq 2r$
$\mathcal{N}(0,\sigma^2 I)$	$l_q(q \leq 2)$	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\big(2G(\frac{\epsilon}{2\sigma}) - 1\big)$	-
	$l_q(q > 2)$	$\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\big(2G(\frac{\epsilon d^{\frac{1}{2}-\frac{1}{q}}}{2\sigma}) - 1\big)$	-

Theorem 4.10. *When smoothing measure is taken as Gaussian probability measure, the certificate $\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\big(2G(\frac{\epsilon}{2\sigma}) - 1\big)$ given in our paper is equivalent to the certificate $\epsilon_{HS,1} \leq [\frac{\theta_a - \theta_b}{2}]_+$ given in paper (Dvijotham et al., 2020a).*

Therefore, when applying both methodologies to Gaussian measure, the formulas obtained are theoretically equivalent. Despite the similarity in analyzing Gaussian measure, our work covers cases with bounded support sets, which is our main contribution.

Then, we discuss the relation with Yang et al. (2020)'s work. As for l_1 adversary, note that for uniform distribution with l_∞ support set, TV-distance-relaxation radius for l_1 adversary has nothing to do with the data dimension d and the certification formula theoretically turns out to be fine, which meets the conclusion in the paper Yang et al. (2020) proves that the Wulff Crystal for the l_1 ball is a cube and the experiment result in figure 1 in their paper that uniform smoothing distribution performs best among the distribution family they list. As for l_∞ adversary, the paper Yang et al. (2020) proves that the Wulff Crystal of l_∞ norm is the zonotope of vectors $\{\pm 1\}^d$, which is a highly complex polytope hard to sample from and indicates the hardness of dealing with l_∞ adversary. And our paper shows that l_∞ adversary magnitude ϵ we can verify must have $\epsilon \leq 2rd^{-1}$ for smoothing measure $\mathcal{U}(B_1(O,r))$ and $\epsilon \leq 2rd^{-\frac{1}{2}}$ for smoothing measure $\mathcal{U}(B_2(O,r))$, which decay at a polynomial rate w.r.t. data dimension d.

5. Experiments

For our adversarial robustness certification task, we choose the *test set certified accuracy* as our metric of interest, which is defined as the fraction of test set that can be correctly classified with a prediction that is also certifiably robust within an l_q ball of an assigned radius r. To pass the robustness certification at data point x, the smoothed classification results of all points within an l_q ball centered at the test point x must be consistent. In our experiment, we mainly focus on l_2 adversary in order to validate part of our theory about the negative certified robustness properties of smoothing measure $\mathcal{U}(B_2(O,r))$ and $\mathcal{U}(B_\infty(O,r))$ in contrast with Gaussian smoothing measure $\mathcal{N}(0,\sigma^2 I)$.

Algorithm 1 Certification Process

```
Input: $T$: test set, target($x$): true class of image $x$, $f(x)$: base classifier, $D(x)$: smoothing distribution, $n$: sample amount, $\epsilon$: perturbation radius, cert(score$_a$, score$_b$, $\epsilon$): certification object
Output: acc: test set certified accuracy
certifiedCount$\leftarrow$ 0,allCount$\leftarrow$ 0
for all $x \in T$ do
    $S \leftarrow \{n$ samples from $D(x)\}$
    count$_c \leftarrow 0$ for every class $c$
    for all $x' \in S$ do
        count$_{f(x')} \leftarrow$ count$_{f(x')} + 1$
    end for
    score$_c \leftarrow$ count$_c$/card$(S)$ for every class $c$
    predict $\leftarrow \arg\max_c\{$score$_c\}$
    if predict = target$(x) \wedge$ cert(score$_c$, $1 -$ score$_c$, $\epsilon$) then
        certifiedCount $\leftarrow$ certifiedCount + 1
    end if
    allCount $\leftarrow$ allCount + 1
end for
return acc$\leftarrow$ certifiedCount/allCount
```

The certification procedure on the test set with assigned l_2 perturbation radius is shown in the following Algorithm 1. Note that the cert(score$_a$, $1 -$ score$_a$, ϵ) function returns true if the certification objective is non-negative, otherwise it returns false, and the objective is calculated using formulas in Table 1 with l_2 perturbation smoothing distribution

$\mathcal{U}(B_2(O,r)), \mathcal{U}(B_\infty(O,r)), \mathcal{N}(0,\sigma^2 I)$. Since our framework directly provides certification objective and involve no iteration and optimization process, our certification procedure only costs constant computation time, which is much faster than Dvijotham et al. (2020a)'s f-divergence-based framework.

We can achieve identical results with Dvijotham et al. (2020a)'s f-divergence-based one w.r.t. Gaussian distribution under specific parameter settings, which is theoretically proved in Section 4.4. However, there is no previous work experimentally examining the properties of uniform smoothing measure, so we mainly focus on comparing Gaussian, l_2, l_∞ ball support set uniform distribution all utilizing our framework.

5.1. Setups

We choose CIFAR-10 as our main dataset and ResNet-110 as our base classifier. We first train the base classifier on the 50000 image training set without smoothing and achieve 89.6% prediction accuracy on the 10000 image test set. Then we run the certification process on the test set with increasing perturbation radius r, test out smoothing distributions as mentioned above and change the parameters σ, r to further illustrate the performance of these distributions. We try increasing the smoothing sample amount to examine the trade-off between computational cost and accuracy improvement. We also run a brief experiment on MNIST dataset to further validate our results, as shown in the left part of Figure 2. The result has similar characteristics with higher general certification rate than the result on CIFAR10 as shown in Figure 3, which will be described in detail in the following section. All training, testing, and certification are run on an NVIDIA RTX 3090.

5.2. Implementation of previous method

Before testing out our own method, we try implementing the previous method introduced by Dvijotham et al. (2020a), which uses iteration method to calculate the certification criteria with Hockey-stick divergence. Since this paper doesn't provide codes, we use an implementation from github (https://github.com/Unispac/F-divergence) as a subject of comparison. We use default settings with Gaussian distribution $N(x, 0.05)$ as smoothing distribution and sample amount n=100, and the optimizing iteration step is set to 20 or 50 to examine its effect on certification accuracy and computational cost.

The result is shown in right part of Figure 2. Our method achieves better certification rate while costing less time. It takes 10mins to run through 10000 samples test dataset with our method while it takes 90 mins for Dvijotham et al. (2020a)'s method with 20 optimizing steps and 3 hours with 50 steps.

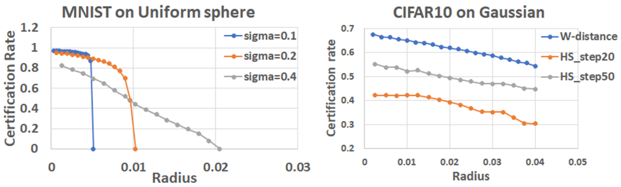

Figure 2. Left: our method's performance on MNIST dataset utilizing l_2 distribution. Right: result of Dvijotham et al. (2020a)'s method with different optimizing steps on CIFAR10 dataset and ours using W-distance.

5.3. Experiments for Different σ, r

We first implement our framework with Gaussian smoothing measure $\mathcal{N}(x, \sigma^2 I)$ where $\sigma = 0.025, 0.05, 0.1$ and sample amount n=100. As shown in Figure 3, there is a neat cut-off for each setting where the perturbation gets too big, and no data can pass the certification at this point. By changing the variance of the smoothing distribution, we observe a clear trend that the increase of variance leads to a drop of initial certification accuracy but also stronger robustness that can endure more significant perturbation; the decrease of the variance leads to the opposite change accordingly.

Next, for smoothing process, we substitute Gaussian distribution with l_2, l_∞ norm ball support set uniform distribution, with $r = 0.025, 0.05, 0.1$. In Figure 3, both experiment results show almost identical characteristics as with Gaussian distribution, but they bring along a critical issue: the mismatch of the perturbation magnitude. Comparing the perturbation radius at the cut-off point, shown in the x-coordinate of Figure 3, we find that the radius of Gaussian distribution is about 50 times larger than that of two uniform distributions.

We assume that this phenomenon is caused by the decreasing intersection of the support sets of smoothing distributions before and after perturbation. For Gaussian distribution, there is always an intersection no matter how large the perturbation radius gets, while for uniform distributions, the support sets will separate quickly and become disjoint under perturbation. Furthermore, the dimension of a 32 × 32 × 3 image is 3072, the square root of which is around 55.4, very close to the cut-off radius's 50 times ratio difference. Such correlation may trace to the involvement of dimension when calculating the finite support set volume of l_2 and l_∞ uniform distribution, while the support set volume of Gaussian distribution is infinitely large. We conjecture that such deficiency is inherent when using the uniform distribution, which can hardly be further improved.

Although methods utilizing such distributions don't achieve satisfying results, our intention to extend the choice of

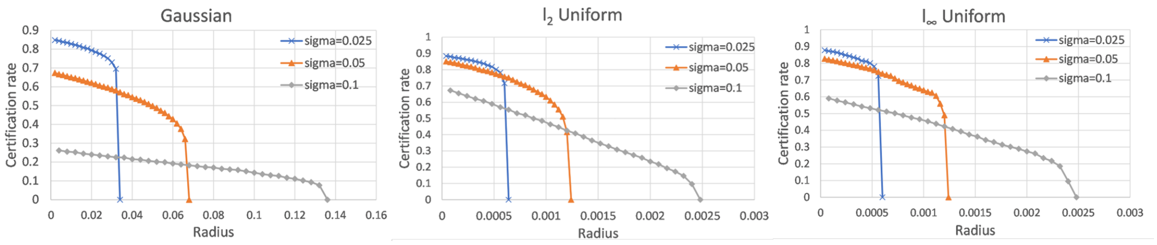

Figure 3. Results of different smoothing distributions using our W-distance and TV-distance based framework. 'Sigma' refers to parameter σ for Gaussian distribution and parameter r for uniform distribution.

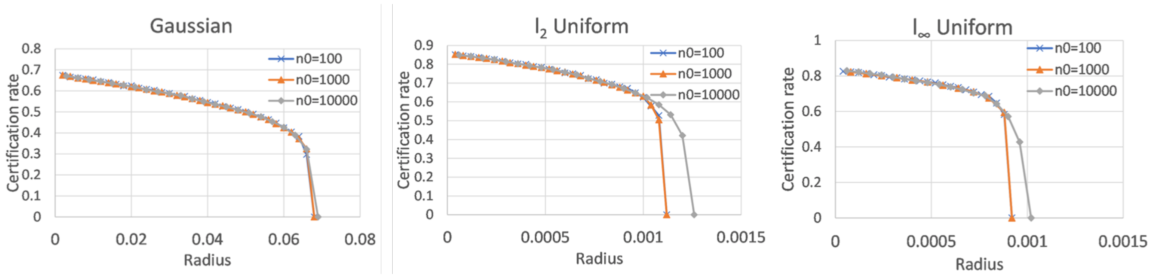

Figure 4. Results on sample amounts with different smoothing distributions using our W-distance and TV-distance based framework.

smoothing distribution is still well realized. Our framework can accommodate more types of smoothing measures, and as shown in the following section, our derivation results can provide great efficiency improvement.

5.4. Experiments for Different Sampling Amounts

When calculating scores for each class in the smoothing process, as we cannot classify all possible data points, we shall only acquire approximate scores by sampling from the smoothed data distribution. Thus such scores may differ in multiple runs due to the randomness of sampling. However, through our experiments, we find that with a certain amount of samples, we can already obtain sufficiently accurate scores, which cannot be significantly improved by increasing the sample amount.

We set the sample amount n to 100, 1000, and 10000 with three different smoothing distributions, and they all obtain similar results: it takes only 10 minutes to run through the 10000 images test set with 100 samples for each image, 30 minutes with 1000 samples and 3 hours with excessive 10000 samples. It is ten times faster than the 3 hours running time with the iteration-based method in Dvijotham et al. (2020a) using just 100 samples. It is also worth noting in Figure 4 that by increasing the sample amount, no significant improvement is observed with Gaussian distribution. However, there is minor progress made with both uniform distributions when the samples are getting overly abundant. We assume that the extra samples make up for the lack of intersections of smoothing uniform distributions before and after the perturbation, while Gaussian distribution has no such issues.

6. Conclusion

We have introduced a framework based on Wasserstein distance and total variation distance relaxation as well as Lagrange duality that is first able to analyze robustness properties of bounded support set smoothing measures both theoretically and experimentally. And by applying our methodology to the simplest bounded support set probability measure example: $\mathcal{U}(B_p(O, r))$, we prove their negative certified robustness properties w.r.t. l_q adversary and present experimental results correspondingly.

Acknowledgments

Yisen Wang is partially supported by the NSF China (No. 62006153), Project 2020BD006 supported by PKU-Baidu Fund, and Open Research Projects of Zhejiang Lab (No. 2022RC0AB05).

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

Balunovic, M. and Vechev, M. Adversarial training and provable defenses: Bridging the gap. In *ICLR*, 2019.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify l_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 2020.

Cao, X. and Gong, N. Z. Mitigating evasion attacks to deep neural networks via region-based classification. In *ACSAC*, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 1982.

Dvijotham, K., Gowal, S., Stanforth, R., Arandjelovic, R., O'Donoghue, B., Uesato, J., and Kohli, P. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018a.

Dvijotham, K., Stanforth, R., Gowal, S., Mann, T. A., and Kohli, P. A dual approach to scalable verification of deep networks. In *UAI*, 2018b.

Dvijotham, K. D., Hayes, J., Balle, B., Kolter, Z., Qin, C., Gyorgy, A., Xiao, K., Gowal, S., and Kohli, P. A framework for robustness certification of smoothed classifiers using f-divergences. In *ICLR*, 2020a.

Dvijotham, K. D., Stanforth, R., Gowal, S., Qin, C., De, S., and Kohli, P. Efficient neural network verification with exactness characterization. In *UAI*. PMLR, 2020b.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of dimensionality on randomized smoothing for certifiable robustness. In *ICML*, 2020.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*. IEEE, 2019.

Lee, G.-H., Yuan, Y., Chang, S., and Jaakkola, T. S. Tight certificates of adversarial robustness for randomly smoothed classifiers. *arXiv preprint arXiv:1906.04948*, 2019.

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

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *ECCV*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *ICML*. PMLR, 2018.

Mirman, M., Singh, G., and Vechev, M. A provable defense for deep residual networks. *arXiv preprint arXiv:1903.12519*, 2019.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Ryou, W., Chen, J., Balunovic, M., Singh, G., Dan, A., and Vechev, M. Fast and effective robustness certification for recurrent neural networks. *arXiv preprint arXiv:2005.13300*, 2020.

Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I., and Bubeck, S. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*, 2019.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. T. Fast and effective robustness certification. *NeurIPS*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thorpe, M. Introduction to optimal transport, 2018.

Wang, S., Chen, Y., Abdou, A., and Jana, S. Mixtrain: Scalable training of verifiably robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018.

Weng, L., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Daniel, L., Boning, D., and Dhillon, I. Towards fast computation of certified robustness for relu networks. In *ICML*. PMLR, 2018.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*. PMLR, 2018.

Wong, E., Schmidt, F. R., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018.

Xiao, K. Y., Tjeng, V., Shafiullah, N. M., and Madry, A. Training for faster adversarial robustness verification via inducing relu stability. *arXiv preprint arXiv:1809.03008*, 2018.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *ICML*. PMLR, 2020.

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. *arXiv preprint arXiv:2001.02378*, 2020.

Zhang, D., Ye, M., Gong, C., Zhu, Z., and Liu, Q. Black-box certification with randomized smoothing: A functional optimization based framework. *Advances in Neural Information Processing Systems*, 33:2316–2326, 2020.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. *arXiv preprint arXiv:1811.00866*, 2018.

A. Notation

We use upper case letters such as K to denote a subset of $\mathbb{R}^d$ and $B_p(O, r)$ to denote a l_p norm ball subset of $\mathbb{R}^d$ centered at original point and with radius r; We use $\mathcal{P}(\mathcal{X})$ to denote the set of all probability measure on measurable space $(\mathcal{X}, \mathcal{F})$ and $\mathcal{D}$ to denote a subset of $\mathcal{P}(\mathcal{X})$, and in our work $\mathcal{X}$ can be seen as a compact set K in $\mathbb{R}^d$ or more generally $\mathbb{R}^d$. We use Greek alphabets μ, ν, ρ to denote probability measure in $\mathcal{P}(\mathcal{X})$; We use $x + \mu$ to denote a new version of probability measure μ with an displacement of $x \in \mathbb{R}^d$; We use upper case letter X to denote a random variable following the Radon-Nikodym derivative of probability measure μ, $x + \mu$, ν or ρ w.r.t. Lebesgue measure λ; We use $W_p(\mu, \nu)$ and $TV(\mu, \nu)$ to denote the W_p distance and total variation distance between probability measures μ and ν; We use $\mathcal{I}$ to denote an indicator function and Vol$(\cdot)$ to denote the volume w.r.t. Lebesgue measure on $\mathbb{R}^d$. And note that throughout the paper, parameter q refer to l_q adversary and parameter p refer to l_p norm ball support set smoothing measure; ϵ refer to the perturbation magnitude of l_q adversary.

B. Optimal Transport Theory

Assume $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Besides, assume μ, ν are absolutely continuous w.r.t. Lebesgue measure λ and let density functions be f and g.

Definition 2 (Push Forward). *If $T : \mathbb{R}^d \to \mathbb{R}^d$, then the distribution of $T(X)$ is called the push-forward of P, denoted by $T_\# P$. In other words,*

$$T_\# P(A) = P(T(x) \in A) = P(T^{-1}(A))$$

Definition 3 (Optimal Distance, Optimal Transport Map). *The Monge version of the optimal transport distance is*

$$\inf_{T : T^\# P = Q} \int \|x - T(x)\|^p dP(x)$$

A minimizer T^*, if one exists, is called the optimal transport map.

Definition 4 (Wasserstein Distance, Earth Mover Distance, Optimal Transport Plan). *Let $\Gamma(\mu, \nu)$ denote all joint distributions γ for (X, Y) that have marginals μ and ν. Then the Wasserstein distance is*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int \|x - y\|_2^p d\gamma(x, y) \right)^{\frac{1}{p}} \text{ where } p \geq 1 \tag{5}$$

When $p = 1$, this is also called the Earth Mover distance. The minimizer γ^* (which does exist) is called the optimal transport plan.

Lemma B.1 (Dual Formulation of Wasserstein Distance When $p \leq 1$). *It can be shown that*

$$W_p^p(\mu, \nu) = \sup_{\psi, \phi} \int \psi(y) d\nu(y) - \int \phi(x) d\mu(x)$$

where $\psi(y) - \phi(x) \leq \|x - y\|^p$. In the special case when $p = 1$, we have the very simple representation

$$\begin{aligned} W_1(\mu, \nu) &= \sup_{\varphi \in \mathcal{F}_1} \int \varphi(x) d\mu(x) - \int \varphi(x) d\nu(x) \\ &= \sup_{\varphi \in \mathcal{F}_1} \int \varphi(x) d(\mu - \nu)(x) = \sup_{\varphi \in \mathcal{F}_1} \int \varphi(x)(f - g)(x) dx \end{aligned} \tag{6}$$

where $\mathcal{F}_1$ denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(x) - f(y)| \leq \|x - y\|$ for all x, y. In the case when $0 < p < 1$, we have similar simple representation

$$\begin{aligned} W_p(\mu, \nu) &= \sup_{\varphi \in \mathcal{F}_p} \int \varphi(x) d\mu(x) - \int \varphi(x) d\nu(x) \\ &= \sup_{\varphi \in \mathcal{F}_p} \int \varphi(x) d(\mu - \nu)(x) = \sup_{\varphi \in \mathcal{F}_p} \int \varphi(x)(f - g)(x) dx \end{aligned} \tag{7}$$

where $\mathcal{F}_p$ denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(x) - f(y)| \leq \|x - y\|^p$ for all x, y.

Lemma B.2 (Dual Formulation of Wasserstein Distance When $1 < p < \infty$)**.** *In the case when $1 < p < \infty$ and the support sets of measure μ and ν are included in a convex compact set K. Define $R = \sup_{x \in K} \|x\|_2$, then we have slightly different dual formulation*

$$\begin{aligned} & W_p(\mu, \nu) \\ \geq & \sup_{\varphi \in Lip(p(2R)^{p-1})} \Big(\int \varphi(y) d(\nu - \mu)(y) - (p-1)(2R)^{p-1} \Big)^{\frac{1}{p}} \\ = & \sup_{\varphi \in Lip(p(2R)^{p-1})} \Big(\int \varphi(y)(g - f)(y) dy - (p-1)(2R)^{p-1} \Big)^{\frac{1}{p}} \end{aligned} \tag{8}$$

where $Lip(p(2R)^{p-1})$ denotes all maps f from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(x) - f(y)| \leq p(2R)^{p-1}\|x - y\|_2$ for all $x, y \in K$.

Definition 5 (Total Variation Distance)**.** *The total variation distance between two probability distribution μ and ν on $\mathbb{R}^d$ is defined by*

$$\|\mu - \nu\|_{TV} = \max\left\{|\mu(A) - \nu(A)| : A \subseteq \mathbb{R}^d\right\}$$

where $\mathbb{R}^d$ is the set of all Borel subsets.

Lemma B.3. *Let μ and ν be two probability distributions on $\mathbb{R}^d$ and absolutely continuous w.r.t. Lebesgue measure λ. Assume the density function of measure μ and ν w.r.t. λ are $f(x)$ and $g(x)$. Then,*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \int_{\mathbb{R}^d} \big|f(x) - g(x)\big| dx$$

C. Proof of Lemma B.2

Recall the dual form of Wasserstein distance

$$W_p^p(\mu, \nu) = \sup_{\psi, \phi \in \mathcal{C}(\mathbb{R}^d)} \int \psi(y) d\nu(y) - \int \phi(x) d\mu(x)$$

where $\psi(y) - \phi(x) \leq \|x - y\|^p$.

For simplicity of the proof, consider equivalent form

$$W_p^p(\mu, \nu) = \sup_{\psi, \phi \in \mathcal{C}(\mathbb{R}^d)} \int \psi(y) d\nu(y) + \int \phi(x) d\mu(x)$$

where $\psi(y) - \phi(x) \leq \|x - y\|^p$. First, we introduce a theorem in (Thorpe, 2018)

Theorem C.1 (Existence of a Maximiser to the Dual Problem)**.** *Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, where X and Y are polish, and $c : X \times Y \to [0, +\infty)$. Assume that there exists $c_X \in L^1(\mu), c_Y \in L^1(\nu)$ such that $c(x, y) \leq c_X(x) + c_Y(y)$ for μ-almost every $x \in X$ and ν-almost every $y \in Y$. In addition, assume that*

$$M := \int_X c_X(x) d\mu(x) + \int_Y c_Y(y) d\nu(y) < \infty$$

Then there exists $(\varphi, \psi) \in \Phi_c = \{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) : \varphi(x) + \psi(y) \leq c(x, y)\}$ where the inequality is understood to hold for μ-almost every $x \in X$ and ν-almost every $y \in Y$ such that

$$\sup_{\Phi_c} \mathbb{J} = \mathbb{J}(\varphi, \psi)$$

where $\mathbb{J}$ is defined by $\mathbb{J} : L^1(\mu) \times L^1(\nu) \to \mathbb{R}, \mathbb{J}(\varphi, \psi) = \int_X \varphi d\mu + \int_Y \psi d\nu$. Futhermore we can choose $(\varphi, \psi) = (\eta^{cc}, \eta^c)$ for some $\eta \in L^1(\mu)$. For $\eta : X \to \bar{\mathbb{R}}$, the c-transforms η^c, η^{cc} are defined by

$$\eta^c : Y \to \bar{\mathbb{R}}, \quad \eta^c(y) = \inf_{x \in X} (c(x, y) - \eta(x))$$

$$\eta^{cc} : Y \to \bar{\mathbb{R}}, \quad \eta^{cc}(y) = \inf_{x \in X} (c(x, y) - \eta^c(x))$$

Lemma C.2. *For $a, b \in \mathbb{R}$ and $1 \leq p < \infty$,*

$$|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$$

Proof. First, it's easy to verify the cases when either of $a = 0, b = 0, p = 1$ holds. Then, Wlog, assume $a, b \in \mathbb{R}^+$

$$\begin{aligned}
&|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p) \\
\iff &(a+b)^p \leq 2^{p-1}(a^p + b^p) \\
\iff &2^{p-1}\Big(\Big(\frac{a}{a+b}\Big)^p + \Big(\frac{b}{a+b}\Big)^p\Big) \geq 1 \\
\iff &2^{p-1}[x^p + (1-x)^p] \geq 1, \forall x \in (0,1)
\end{aligned}$$

where the last inequality is easy to verify. □

In our case, $c(x,y) = \|x-y\|^p \leq (\|x\| + \|y\|)^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$ and the requirement that $M < \infty$ is exactly the condition that μ and ν have finite p^{th} moments which is easy to verify by noting that supp$(\mu) =$ supp$(\nu) = K$ is compact set in $\mathbb{R}^d$. Then, according to the theorem, there exists $\eta \in L^1(\mu)$ such that

$$W_p^p(\mu,\nu) = \sup_{\eta \in L^1(\mu)} \int \eta^c(y) d\nu(y) + \int \eta^{cc}(x) d\mu(x)$$

Note that η^c possesses Lipschitz continuous property stated below

Lemma C.3. *For $\eta \in L^1(K)$ where $K \subseteq \mathbb{R}^d$ is a convex compact set, then η^{c_p} is a $p(2R)^{p-1}$-Lipschitz function where $R := \sup_{x \in K} \|x\|$ and $c_p(x,y) = \|x-y\|^p$, i.e.,*

$$\|\eta^{c_p}(x) - \eta^{c_p}(y)\| \leq p(2R)^{p-1}\|x-y\|, \quad x, y \in K$$

Proof.

$$\begin{aligned}
&|\eta^{c_p}(x) - \eta^{c_p}(y)| \\
=&\Big|\inf_{z_1 \in K}\big(\|x-z_1\|^p - \eta(z_1)\big) - \inf_{z_2 \in K}\big(\|y-z_2\|^p - \eta(z_2)\big)\Big| \qquad (9) \\
=&\Big|\inf_{z_1 \in K}\sup_{z_2 \in K}\Big(\big(\|x-z_1\|^p - \|y-z_2\|^p\big) - \big(\eta(z_1) - \eta(z_2)\big)\Big)\Big| \\
\leq &\sup_{z \in K}\Big|\big(\|x-z\|^p - \eta(z)\big) - \big(\|y-z\|^p - \eta(z)\big)\Big| \qquad (10) \\
= &\sup_{z \in K}\Big|\|x-z\|^p - \|y-z\|^p\Big|
\end{aligned}$$

where 9 is due to the definition of c-transform; 10 is obtained by taking a specific value of z_1 as z_2. Note that K is a compact set and $\big|\|x-z\|^p - \|y-z\|^p\big|$ is a continuous function w.r.t. z, then there exists a point z^* such that $\big|\|x-z^*\|^p - \|y-z^*\|^p\big| = \sup_{z \in K}\big|\|x-z\|^p - \|y-z\|^p\big|$. According to the first order condition, z^* satisfies the equation below

$$
\begin{align}
&\nabla_z(\|x-z\|^p - \|y-z\|^p) \notag\\
=&\nabla_z\|x-z\|^p - \nabla_z\|y-z\|^p \notag\\
=&\nabla_z\|z-x\|^p - \nabla_z\|z-x\|^p \notag\\
=&p\|z-x\|^{\frac{p}{2}-1}(z-x)^\top - p\|z-y\|^{\frac{p}{2}-1}(z-y)^\top = 0 \tag{11}\\
\Longrightarrow&\|z-x\|^{\frac{p}{2}-1}(z-x)^\top = \|z-y\|^{\frac{p}{2}-1}(z-y)^\top \notag\\
\Longrightarrow&(\|z-x\|^{\frac{p}{2}-1} - \|z-y\|^{\frac{p}{2}-1})z^\top \notag\\
&= \|z-x\|^{\frac{p}{2}-1}x^\top - \|z-y\|^{\frac{p}{2}-1}y^\top \notag\\
\Longrightarrow&z = \frac{\|z-x\|^{\frac{p}{2}-1}}{\|z-x\|^{\frac{p}{2}-1} - \|z-y\|^{\frac{p}{2}-1}}x \notag\\
&- \frac{\|z-y\|^{\frac{p}{2}-1}}{\|z-x\|^{\frac{p}{2}-1} - \|z-y\|^{\frac{p}{2}-1}}y \notag
\end{align}
$$

where 11 is due to $\nabla_x\|x\|^p = \nabla_x(x^\top x)^{\frac{p}{2}} = p(x^\top x)^{\frac{p}{2}-1}x^\top = p\|x\|^{\frac{p}{2}-1}x^\top$. And this equation shows that z^* lie on the line determined by x and y but does not lies on the part between x and y, which can be formulated as $z^* = \lambda x + (1-\lambda)y, \lambda \in \mathbb{R} \setminus (0,1)$. Note that

$$
\begin{aligned}
&\sup_{\lambda\in\mathbb{R}\setminus(0,1)} \Big\| |x-z^*\|^p - \|y-z^*\|^p \Big| \\
=&\sup_{\lambda\in\mathbb{R}\setminus(0,1)} \Big\| |(1-\lambda)(x-y)\|^p - \|\lambda(y-x)\|^p \Big| \\
=&\sup_{\lambda\in\mathbb{R}\setminus(0,1)} \Big\| 1-\lambda|^p - |\lambda|^p \Big| \cdot \|x-y\|^p \\
&= \Big(\sup_{\lambda\in\mathbb{R}\setminus(0,1)} \Big\| 1-\lambda\|^p - |\lambda|^p \Big| \Big) \cdot \|x-y\|^p
\end{aligned}
$$

Then, we just need to optimize

$$
\begin{aligned}
&\sup_{\lambda\in\mathbb{R}\setminus(0,1)} \Big\| 1-\lambda|^p - |\lambda|^p \Big| \\
&s.t. \quad \lambda x + (1-\lambda)y \in K
\end{aligned}
$$

Note that we can relax the constraint as below

$$
\begin{align}
&\lambda x + (1-\lambda)y \in K \notag\\
\Longleftrightarrow\ &\lambda(x-y) + y = (1-\lambda)(y-x) + x \in K \notag\\
\Longrightarrow&\|\lambda(x-y)+y\| = \|(1-\lambda)(y-x)+x\| \le R \tag{12}\\
\Longrightarrow&\|\lambda(x-y)\| \le R + \|y\|, \notag\\
&\qquad \|(1-\lambda)(y-x)\| \le R + \|x\| \tag{13}\\
\Longrightarrow&|\lambda| \cdot \|x-y\| \le 2R, \quad |1-\lambda| \cdot \|x-y\| \le 2R \tag{14}\\
\Longrightarrow&1 - \frac{2R}{\|x-y\|} \le \lambda \le \frac{2R}{\|x-y\|} \notag
\end{align}
$$

where 12 and 14 is due to the definition of R as $\sup_{x\in K}\|x\|$; 13 is due to triangular inequality.

Using the relaxed constraint, we can show that when $\lambda \ge 1$, $\Big\| 1-\lambda|^p - |\lambda|^p \Big| = \lambda^p - (\lambda-1)^p$ is an increasing function w.r.t. λ as $p \ge 1$, then

$$
\begin{aligned}
\Big\| 1-\lambda|^p - |\lambda|^p \Big| &= \lambda^p - (\lambda-1)^p \\
&\le \left(\frac{2R}{\|x-y\|}\right)^p - \left(\frac{2R}{\|x-y\|} - 1\right)^p
\end{aligned}
$$

And when $\lambda \leq 0$, $\Big||1-\lambda|^p - |\lambda|^p\Big| = (1-\lambda)^p - (-\lambda)^p$ is a decreasing function w.r.t λ as $p \geq 1$, then

$$\begin{aligned}\Big||1-\lambda|^p - |\lambda|^p\Big| &= (1-\lambda)^p - (-\lambda)^p \\ &\leq \left(\frac{2R}{\|x-y\|}\right)^p - \left(\frac{2R}{\|x-y\|} - 1\right)^p\end{aligned}$$

Note that

$$\begin{aligned}&\left(\frac{2R}{\|x-y\|}\right)^p - \left(\frac{2R}{\|x-y\|} - 1\right)^p \\ =&p\bigg(k \cdot \Big(\frac{2R}{\|x-y\|}\Big) + (1-k)\cdot\Big(\frac{2R}{\|x-y\|} - 1\Big)\bigg)^{p-1} \\ =&p\bigg(\frac{2R}{\|x-y\|} + (k-1)\bigg)^{p-1} \leq p\bigg(\frac{2R}{\|x-y\|}\bigg)^{p-1}\end{aligned} \tag{15}$$

where 15 is due to the Differential Mean Value Theorem where $k \in (0,1)$.

Thus, we have

$$\begin{aligned}|\eta^{c_p}(x) - \eta^{c_p}(y)| &\leq p\bigg(\frac{2R}{\|x-y\|}\bigg)^{p-1} \cdot \|x-y\|^p \\ &= p(2R)^{p-1}\|x-y\|\end{aligned}$$

i.e. $\eta^{c_p}(x)$ is a $p(2R)^{p-1}$-Lipschitz function. □

Using Lipschitz continuous property of η^c, we get

$$W_p^p(\mu,\nu) = \sup_{\eta \in L^1(\mu)} \int \eta^c(y)d\nu(y) + \int \eta^{cc}(x)d\mu(x) \tag{16}$$

$$\leq \sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y)d\nu(y) + \int \varphi^c(x)d\mu(x)$$

$$= \sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y)d\nu(y) + \int \varphi^c(y)d\mu(y) \tag{17}$$

where $\mathbf{Lip}(p(2R)^{p-1})$ denotes the set of $p(2R)^{p-1}$-Lipschitz functions. On the other hand, recall that

$$W_p^p(\mu,\nu) = \sup_{\psi,\phi \in \mathcal{C}(\mathbb{R}^d)} \int \psi(y)d\nu(y) + \int \phi(x)d\mu(x)$$

where $\psi(y) + \phi(x) \leq \|x-y\|^p$. Keeping $\psi(x)$ fixed and optimizing w.r.t. $\phi(y)$, then we just need to optimize $\int \phi(y)d\mu(y)$ under constraint $\phi(y) \leq \|x-y\|^p - \psi(x)$. Then obviously we have $\phi^*(y) = \inf_{x \in K}\big(\|x-y\|^p - \psi(x)\big) = \psi^{c_p}(y)$ where $c_p(x,y) = \|x-y\|^p$. The map $(\phi,\psi) \in \mathcal{C}(K)^2 \mapsto (\psi^{c_p},\psi) \in \mathcal{C}(K)^2$ replaces dual potentials by ”better” ones improving the dual objective $W_p^p(\mu,\nu)$.

Using c-transform, we can reformulate constrained problem into unconstrained convex problem over a single potential

$$\begin{aligned}W_p^p(\mu,\nu) &= \sup_{\psi \in \mathcal{C}(\mathbb{R}^d)} \int \psi(y)d\nu(y) + \int \psi^{c_p}(x)d\mu(x) \\ &= \sup_{\psi \in \mathcal{C}(\mathbb{R}^d)} \int \psi(y)d\nu(y) + \int \psi^{c_p}(y)d\mu(y)\end{aligned} \tag{18}$$

Combining 17 and 18, we know that when the support set of measure μ and ν supp(μ) = supp$(\nu) = K$ where K is a convex compact set, we have

$$\begin{aligned}&\sup_{\psi \in \mathcal{C}(K)} \int \psi(y)d\nu(y) + \int \psi^{c_p}(y)d\mu(y) = W_p^p(\mu,\nu) \\ \leq &\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y)d\nu(y) + \int \varphi^c(y)d\mu(y)\end{aligned} \tag{19}$$

Note that Lipschitz function must be continuous and therefore **Lip**$(p(2R)^{p-1}) \subseteq \mathcal{C}(K)$. Then, we have

$$\begin{aligned}
&\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d\nu(y) + \int \varphi^c(y) d\mu(y) \\
\leq &\sup_{\phi \in \mathcal{C}(K)} \int \psi(y) d\nu(y) + \int \psi^{c_p}(y) d\mu(y)
\end{aligned} \tag{20}$$

Combining 19 and 20, we know the inequality in 19 changes into equality

$$W_p^p(\mu, \nu) = \sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d\nu(y) + \int \varphi^c(y) d\mu(y) \tag{21}$$

Note that for $f(x) = x^p - p(2R)^{p-1}x, x \in \mathbb{R}^+$ achieves its minimum when $f'(x) = px^{p-1} - p(2R)^{p-1} = 0$, i.e. $x = 2R$ and the minimum is $f(2R) = -(p-1)(2R)^{p-1}$. Then,

$$\begin{aligned}
\varphi^{c_p}(y) &= \inf_{x \in K} \Big(\|x - y\|^p - \varphi(x) \Big) \\
&\geq \inf_{x \in K} \Big(\|x - y\|^p - \varphi(y) - p(2R)^{p-1} \|x - y\| \Big) \\
&= -\varphi(y) - (p-1)(2R)^{p-1}
\end{aligned} \tag{22}$$

Thus, we attain a lower bound of $W_p^p(\mu, \nu)$

$$\begin{aligned}
&W_p^p(\mu, \nu) \\
= &\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d\nu(y) + \int \varphi^c(y) d\mu(y) && (23) \\
\geq &\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d\nu(y) - \int \Big(\varphi(y) + (p-1)(2R)^{p-1} \Big) d\mu(y) && (24) \\
= &\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d\nu(y) - \int \varphi(y) d\mu(y) - (p-1)(2R)^{p-1} \\
= &\sup_{\varphi \in \mathbf{Lip}(p(2R)^{p-1})} \int \varphi(y) d(\nu - \mu)(y) - (p-1)(2R)^{p-1}
\end{aligned}$$

where 23 is due to 21 and 24 is due to 22.

D. Proof of Theorem 4.1

Proof.

$$\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \leq \epsilon\} \tag{25}$$

Note that

$$\begin{aligned}\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} W_p(\mu, \nu) &= \sup_{\|x-x'\|_q \leq \epsilon} W_p(x+\mu, x'+\mu) \\ &= \sup_{\|z\|_q \leq \epsilon} W_p(\mu, z+\mu)\end{aligned}$$

where the first equality is due to the definition of $\mathcal{D}_{x,\epsilon,q}$ and the second equality is due to the translation invariance property of Wasserstein distance.

Then recall the Monge version of Wasserstein distance

$$W_p(\mu,\nu) \leq \left(\inf_{T:T^{\#}\mu=\nu} \int \|x - T(x)\|^p d\mu(x)\right)^{\frac{1}{p}}$$

Noticing the **inf** operator in the Monge version definition of W_p, we can get an upper bound for $W_p(\mu,\nu)$ by specializing a transport map $\tilde{T}$ satisfying $\tilde{T}\mu = \nu$. In our case, we take $\tilde{T} : \mathbb{R}^d \to \mathbb{R}^d, \tilde{T} : x \mapsto x + z$, and it's easy to verify that $\tilde{T}^{\#}\mu = z + \mu$. Then we get the upper bound below

$$\begin{aligned}W_p(\mu, z+\mu) \leq& \left(\inf_{T:T^{\#}\mu=z+\mu} \int \|x - T(x)\|^p d\mu(x)\right)^{\frac{1}{p}} \\ \leq& \left(\int \|x - \tilde{T}(x)\|^p d\mu(x)\right)^{\frac{1}{p}} = \|z\|\end{aligned}$$

where the last equality is due to μ is a probability measure. This provides us with an intuition that the upper bound of $W_p(\mu, z+\mu)$ is determined by the Euclidean norm of displacement z. Using this upper bound,

$$\sup_{\|z\|_q \leq \epsilon} W_p(\mu, z+\mu) \leq \sup_{\|z\|_q \leq \epsilon} \|z\|_2$$

When $0 < q \leq 2$, using the lemma that when $0 < p < q < \infty$, $\|z\|_q \leq \|z\|_p, \forall z \in \mathbb{R}^d$ holds, we have $\sup_{\|z\|_q \leq \epsilon} \|z\|_2 \leq \sup_{\|z\|_q \leq \epsilon} \|z\|_q = \epsilon$. On the other hand, note that $\|\epsilon e_1\|_2 = \|\epsilon e_1\|_q = \epsilon$, we have $\sup_{\|z\|_q \leq \epsilon} \|z\|_2 = \epsilon$. And when $q > 2$, recall Holder's Inequality below

Lemma D.1 (Holder's Inequality for $\mathbb{R}^n$)**.** *For* $\{a_i\}_{1 \leq i \leq n}, \{b_i\}_{1 \leq i \leq n} \subseteq \mathbb{R}, r > 1$, *we have*

$$\sum_{i=1}^{n} |a_i||b_i| \leq \left(\sum_{i=1}^{n} |a_i|^r\right)^{\frac{1}{r}} \left(\sum_{i=1}^{n} |a_i|^{\frac{r}{r-1}}\right)^{\frac{r-1}{r}}$$

Apply it to the case $n = d, |a_i| = |x_i|^2, |b_i| = 1$ and $r = \frac{q}{2} > 1$,

$$\begin{aligned}&\sum_{i=1}^{d} |x_i|^2 = \sum_{i=1}^{d} |x_i|^2 \cdot 1 \\ \leq& \left(\sum_{i=1}^{d} (|x_i|^2)^{\frac{q}{2}}\right)^{\frac{2}{q}} \left(\sum_{i=1}^{d} 1^{\frac{q}{q-2}}\right)^{\frac{q-2}{q}} = \left(\sum_{i=1}^{d} |x_i|^q\right)^{\frac{2}{q}} d^{1-\frac{2}{q}}\end{aligned}$$

$$\begin{aligned}\|x\|_2 =& \left(\sum_{i=1}^{d} |x_i|^2\right)^{\frac{1}{2}} \\ \leq& \left(\sum_{i=1}^{d} |x_i|^q\right)^{\frac{1}{q}} d^{\frac{1}{2}-\frac{1}{q}} = \|x\|_q d^{\frac{1}{2}-\frac{1}{q}}\end{aligned}$$

Thus,

$$\sup_{\|z\|_q \leq \epsilon} \|z\|_2 \leq \sup_{\|z\|_q \leq \epsilon} \|x\|_q d^{\frac{1}{2}-\frac{1}{q}} = \epsilon d^{\frac{1}{2}-\frac{1}{q}}.$$

On the other hand, note that

$$\Big\|\frac{\epsilon}{n^{\frac{1}{q}}} \sum_{i=1}^{d} e_i\Big\|_q = \epsilon, \ \Big\|\frac{\epsilon}{n^{\frac{1}{q}}} \sum_{i=1}^{d} e_i\Big\|_2 = \epsilon d^{\frac{1}{2}-\frac{1}{q}},$$

we have

$$\sup_{\|z\|_q \leq \epsilon} \|z\|_2 = \epsilon d^{\frac{1}{2}-\frac{1}{q}}.$$

Combining the case when $0 < q \leq 2$ and $q > 2$, we have

$$\sup_{\|x-x'\|_q \leq \epsilon} W_p(x+\mu, x'+\mu) = \sup_{\|z\|_q \leq \epsilon} W_p(\mu, z+\mu) \leq \begin{cases} \epsilon \text{ when } 0 < q \leq 2 \\ \epsilon d^{\frac{1}{2}-\frac{1}{q}} \text{ when } q > 2 \end{cases} = \max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}$$

□

E. W_2 distance relaxation is tight for Gaussian Probability Measure

Here, we show that W_2 distance relaxation for Gaussian probability measure is tight.

Theorem E.1. *When $\mu = \mathcal{N}(0, \sigma^2 I)$ and $p = 2$, the relaxation in* **??** *is tight. In other words,*

$$\begin{aligned}&\mathcal{D}_{x,\epsilon,q} \subseteq \mathcal{D}_{x,\max\{\epsilon,\epsilon d^{\frac{1}{2}-\frac{1}{q}}\},2}\\ &\textit{but } \mathcal{D}_{x,\epsilon,q} \setminus \mathcal{D}_{x,\max\{\epsilon,\epsilon d^{\frac{1}{2}-\frac{1}{q}}\}-\delta,2} \neq \emptyset\\ &\textit{for any sufficiently small } \delta > 0.\end{aligned} \tag{26}$$

Proof. Note that (Dowson & Landau, 1982) established the formula of Wasserstein distance between two Gaussian measures.

Theorem E.2. *For Gaussian probability measures $\mu = \mathcal{N}(\mu_1, \Sigma_1)$ and $\nu = \mathcal{N}(\mu_2, \Sigma_2)$, W_2-distance between μ and ν have closed form formula*

$$W_2(\mu, \nu)^2 = \|\mu_1 - \mu_2\|^2 + \mathit{tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}\big) \tag{27}$$

Using above theorem, we yield following tight relaxation between norm-based constraint set $\mathcal{D}_{x,\epsilon,q}$ and W_2-distance based constraint sets $\mathcal{D}_{x,\delta,2}$ for Gaussian smoothing measures centered at origin, i.e. $\mu = \mathcal{N}(0, \sigma^2 I)$

$$\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} W_2(\mu, \nu) = \sup_{\|x-x'\|_q \le \epsilon} W_2(x + \mu, x' + \mu) \tag{28}$$

$$= \sup_{\|z\|_q \le \epsilon} W_2(\mu, z + \mu)$$

$$= \sup_{\|z\|_q \le \epsilon} \|z\|_2 = \max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\} \tag{29}$$

where 29 is due to theorem E.2 and equality D. And generalization of above theorem when $\mu = \mathcal{N}(0, \Sigma)$ can be proved in the same way. □

F. Proof of Theorem 4.2

Proof. First, we introduce the lemma below.

Lemma F.1. *Let X be a random variable that follows d-dimensional Gaussian distribution with density function*

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

where $x, \mu \in \mathbb{R}^d$ and $\Sigma \in S^d_{++}$. Let $H : x^T w + b = 0$ be a hyperplane in the d-dimensional Euclidean space $\mathbb{R}^d$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The hyperplane H defines two half-spaces:

$$\begin{aligned}\Omega_+ &= \{x \in \mathbb{R}^d | x^T w + b \geq 0\}, \\ \Omega_- &= \{x \in \mathbb{R}^d | x^T w + b < 0\}\end{aligned}$$

Define the integral over half-space Ω_+ as

$$\begin{aligned}P &= \int_{\Omega_+} f(x; \mu, \Sigma) dx \\ &= \int_{\Omega_+} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} dx\end{aligned}$$

Since Σ is positive definite symmetric, there exist an orthogonal matrix U and a diagonal matrix D with positive diagonal elements such that $\Sigma = U^T D U$. Let $x_0 = -\frac{\mu^\top w + b}{\|\sqrt{D} U w\|_2}$ and hence $P = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$.

(The proof of this lemma is credit to https://math.stackexchange.com/questions/556977/gaussian-integrals-over-a-half-space.)

Recall the definition of l_p-norm constraint set of probability measures

$$\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \leq \epsilon\}$$

Note that

$$\begin{aligned}\sup_{\nu \in \mathbb{D}_{x,\epsilon,q}} TV(\mu, \nu) &= \sup_{\|x - x'\|_q \leq \epsilon} TV(x + \mu, x' + \mu) \\ &= \sup_{\|z\|_q \leq \epsilon} TV(\mu, z + \mu)\end{aligned}$$

where the first equality is due to the definition of $\mathcal{D}_{x,\epsilon,q}$ and the second equality is due to the translation invariance property of total variation distance.

Define hyperplane $H^1 : x^T z - \frac{\|z\|_2^2}{2} = 0$ and $H^2 : x^T z + \frac{\|z\|_2^2}{2} = 0$. The hyperplane H^1 defines two half-spaces: $\Omega^1_+ = \{x \in \mathbb{R}^d | x^T z - \frac{\|z\|_2^2}{2} \geq 0\}$ and $\Omega^1_- = \{x \in \mathbb{R}^d | x^T z - \frac{\|z\|_2^2}{2} < 0\}$. And the hyperplane H^2 defines two half-spaces: $\Omega^2_+ = \{x \in \mathbb{R}^d | x^T z + \frac{\|z\|_2^2}{2} \geq 0\}$ and $\Omega^2_- = \{x \in \mathbb{R}^d | x^T z + \frac{\|z\|_2^2}{2} < 0\}$. Applying lemma F.1 and lemma B.3, we know that

$$
\begin{aligned}
&\sup_{\|z\|_q \leq \epsilon} TV(\mu, z+\mu) \\
&= \sup_{\|z\|_q \leq \epsilon} \frac{1}{2} \int \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \left| e^{-\frac{x^T x}{2\sigma^2}} - e^{-\frac{(x-z)^T (x-z)}{2\sigma^2}} \right| dx && (30) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \int_{\Omega_+^1} \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \left| e^{-\frac{x^T x}{2\sigma^2}} - e^{-\frac{(x-z)^T (x-z)}{2\sigma^2}} \right| dx \\
&\qquad + \int_{\Omega_-^1} \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \left| e^{-\frac{x^T x}{2\sigma^2}} - e^{-\frac{(x-z)^T (x-z)}{2\sigma^2}} \right| dx && (31) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \int_{\Omega_+^1} \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \left(e^{-\frac{(x-z)^T (x-z)}{2\sigma^2}} - e^{-\frac{x^T x}{2\sigma^2}} \right) dx \\
&\qquad\quad + \int_{\Omega_-^1} \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \left(e^{-\frac{x^T x}{2\sigma^2}} - e^{-\frac{(x-z)^T (x-z)}{2\sigma^2}} \right) dx && (32) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \int_{\Omega_+^1} d(z+\mu) - \int_{\Omega_+^1} d\mu + \int_{\Omega_-^1} d\mu - \int_{\Omega_-^1} d(z+\mu) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \int_{\Omega_+^2} d\mu - \int_{\Omega_+^1} d\mu + \int_{\Omega_-^1} d\mu - \int_{\Omega_-^2} d\mu && (33) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \int_{\frac{-\|z\|_2}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx - \int_{\frac{\|z\|_2}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\
&\qquad\quad + \int_{-\infty}^{\frac{\|z\|_2}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx - \int_{-\infty}^{-\frac{\|z\|_2}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx && (34) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} \Bigg(G\Big(\frac{\|z\|_2}{2\sigma}\Big) - G\Big(-\frac{\|z\|_2}{2\sigma}\Big) \\
&\qquad\qquad\qquad + G\Big(\frac{\|z\|_2}{2\sigma}\Big) - G\Big(-\frac{\|z\|_2}{2\sigma}\Big)\Bigg) && (35) \\
&= \frac{1}{2} \sup_{\|z\|_q \leq \epsilon} 2\Big(2G\Big(\frac{\|z\|_2}{2\sigma}\Big) - 1\Big) && (36) \\
&= 2G\Big(\frac{\max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}}{2\sigma}\Big) - 1 && (37)
\end{aligned}
$$

where 30 is due to lemma B.3; 32 is due to the consistency of sign of integrand function on Ω_+^1 and Ω_-^1; 33 is due to the transformation formula of space coordinates; 34 is due to lemma F.1; 35 and 37 is due to the definition and central symmetry property of G as the cumulative density function of standard normal distribution; 37 is due to D. □

G. Proof of Theorem 4.3

Proof. Recall the definition of l_p-norm constraint set of probability measures

$$\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \leq \epsilon\}$$

Note that

$$\begin{aligned}\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu,\nu) &= \sup_{\|x-x'\|_q \leq \epsilon} TV(x+\mu, x'+\mu) \\ &= \sup_{\|z\|_q \leq \epsilon} TV(\mu, z+\mu)\end{aligned}$$

where the first equality is due to the definition of $\mathcal{D}_{x,\epsilon,q}$ and the second equality is due to the translation invariance property of total variation distance. Next, compute the value of $TV(\mu, z+\mu)$.

Lemma G.1. *K is a l_1 norm ball centered at original point of radius r, then $K \cap (z+K) = \emptyset$ if and only if $\|z\|_1 > 2r$.*

Proof. First, we prove the if part and assume $\|z\|_1 > 2r$. Consider arbitrarily taken $x \in (z+K)$, i.e. $\|x-z\|_1 \leq r$. According to the triangular inequality with respect to l_1 norm, we have

$$\|x\|_1 = \|z - (x-z)\|_1 \geq \|z\|_1 - \|x-z\|_1 > 2r - r = r \tag{38}$$

which shows that $x \notin K$ and therefore $K \cap (z+K) = \emptyset$.
Then we prove the only if part by using reduction to absurdity and assume $\|z\|_1 \leq 2r$. Take $y = \frac{1}{2}z$, then $\|y\|_1 = \frac{1}{2}\|z\|_1 \leq r$ and $\|y - z\|_1 = \frac{1}{2}\|z\|_1 \leq r$ which shows that $y \in K \cap (z+K)$ and therefore $K \cap (z+K) \neq \emptyset$ which leads to a contradiction. □

According to lemma G.1, we know that when $\left\{z \middle| \|z\|_q \leq \epsilon, \|z\|_1 \geq 2r\right\} \neq \emptyset$, we have

$$\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu,\nu) = \sup_{\|z\|_q \leq \epsilon} TV(\mu, z+\mu) = 1$$

Define $\bar{z} = \frac{2r}{d}\sum_{i=1}^{d} e_i$, and it's easy to verify that $\|\bar{z}\|_1 = 2r$ and $\|\bar{z}\|_q = 2rd^{\frac{1}{q}-1}$ for $q > 1$. Thus, when $\epsilon > 2rd^{\frac{1}{q}-1}$, we have

$$\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu,\nu) = \sup_{\|z\|_q \leq \epsilon} TV(\mu, z+\mu) = 1$$

□

H. Proof of Theorem 4.4

Proof. First, we introduce the lemmas below for the convenience of later proof.

Lemma H.1 (Volume Formula of d-dimensional spherical cap)**.** *The volume of a d-dimensional hyperspherical cap of height h and radius r is given by:*

$$V = \frac{\pi^{\frac{d-1}{2}} r^d}{\Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{r-h}{r})} \sin^d(t)dt$$

where we define h as the value shown in figure 5 and Γ (the gamma function) is given by $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$.

Lemma H.2 (Volume formula of d-dimensional Euclidean ball)**.** *The volume of d-dimensional Euclidean ball of radius r is given by*

$$V = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2}+1)}$$

Recall the definition of l_p-norm constraint set of probability measures

$$\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \leq \epsilon\}$$

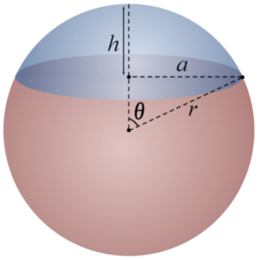

Figure 5. An example of a spherical cap in blue

Note that

$$\begin{aligned}\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu,\nu) &= \sup_{\|x-x'\|_q \le \epsilon} TV(x+\mu, x'+\mu) \\ &= \sup_{\|z\|_q \le \epsilon} TV(\mu, z+\mu)\end{aligned}$$

where the first equality is due to the definition of $\mathcal{D}_{x,\epsilon,q}$ and the second equality is due to the translation invariance property of total variation distance.

Lemma H.3. *K is a l_2 norm ball centered at original point of radius r, then $K \cap (z+K) = \emptyset$ if and only if $\|z\|_2 > 2r$.*

According to this lemma, we know that when $q \le 2$ and $\epsilon > 2r$, we have

$$\begin{aligned}1 \ge &\sup_{\|z\|_q \le \epsilon} TV(\mu, z+\mu) \\ &\ge TV(\mu, \epsilon e_1 + \mu) = \frac{\text{Vol}(K\Delta(\epsilon e_1 + K))}{2\text{Vol}(K)} = 1\end{aligned}$$

where the last equality is due to $\|\epsilon e_1\|_2 = \epsilon > 2r$ and applying lemma H.3. And when $q > 2$ and $\epsilon > 2rd^{\frac{1}{q}-\frac{1}{2}}$, we have

$$\begin{aligned}1 \ge &\sup_{\|z\|_q \le \epsilon} TV(\mu, z+\mu) \ge TV(\mu, \frac{\epsilon}{d^{\frac{1}{q}}}\sum_{i=1}^{d} e_i + \mu) \\ =&\frac{\text{Vol}\Big(K\Delta\big(\frac{\epsilon}{d^{\frac{1}{q}}}\sum_{i=1}^{d} e_i + K\big)\Big)}{2\text{Vol}(K)} = 1\end{aligned}$$

where the last equality is due to $\|\frac{\epsilon}{d^{\frac{1}{q}}}\sum_{i=1}^{d} e_i\|_2 = \epsilon d^{\frac{1}{2}-\frac{1}{q}} > 2r$ and applying lemma H.3. Combining the results for $q \le 2$ and $q > 2$, we have

$$\sup_{\|z\|_q \le \epsilon} TV(\mu, z+\mu) = 1 \text{ when } \epsilon > \min\{2r, 2rd^{\frac{1}{q}-\frac{1}{2}}\}$$

Next, consider the case when $\epsilon \leq \min\{2r, 2rd^{\frac{1}{q}-\frac{1}{2}}\}$. Applying H.1, lemma H.2 and lemma B.3, we have

$$\begin{aligned}
&\sup_{\|z\|_q \leq \epsilon} TV(\mu, z+\mu) \\
=& \sup_{\|z\|_q \leq \epsilon} \frac{1}{2} \int \left| \frac{1}{\text{Vol}(K)} \mathcal{I}_{x \in K} - \frac{1}{\text{Vol}(K)} \mathcal{I}_{x \in z+K} \right| dx \qquad (39) \\
=& \sup_{\|z\|_q \leq \epsilon} \frac{1}{2\text{Vol}(K)} \int \mathcal{I}_{x \in K \Delta (z+K)} dx \\
=& \sup_{\|z\|_q \leq \epsilon} \frac{\text{Vol}(K \Delta (z+K))}{2\text{Vol}(K)} \\
=& \sup_{\|z\|_q \leq \epsilon} \frac{\text{Vol}(K) - \frac{2\pi^{\frac{d-1}{2}} r^d}{\Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{\|z\|_2}{2r})} \sin^d(t) dt}{\text{Vol}(K)} \\
=& \sup_{\|z\|_q \leq \epsilon} 1 - \frac{\frac{2\pi^{\frac{d-1}{2}} r^d}{\Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{\|z\|_2}{2r})} \sin^d(t) dt}{\text{Vol}(K)} \qquad (40) \\
=& \sup_{\|z\|_q \leq \epsilon} 1 - \frac{\frac{2\pi^{\frac{d-1}{2}} r^d}{\Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{\|z\|_2}{2r})} \sin^d(t) dt}{\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d} \\
=& \sup_{\|z\|_q \leq \epsilon} 1 - \frac{2\Gamma(\frac{d}{2}+1)}{\pi^{\frac{1}{2}} \Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{\|z\|_2}{2r})} \sin^d(t) dt \qquad (41) \\
=& 1 - \frac{2\Gamma(\frac{d}{2}+1)}{\pi^{\frac{1}{2}} \Gamma(\frac{d+1}{2})} \int_0^{\arccos(\frac{\max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}}{2r})} \sin^d(t) dt \qquad (42)
\end{aligned}$$

where 39 is due to lemma B.3; 40 is due to lemma H.1; 41 is due to lemma H.2; 42 is due to D. Because of the computation difficulty (overflow), we have to simplify the term $\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})}$.

When d is even, assume $d = 2k, k \in \mathbb{N}$ and note that $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$, then

$$\begin{aligned}
\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} =& \frac{\Gamma(k+1)}{\Gamma(k+\frac{1}{2})} = \frac{k!\Gamma(1)}{\Pi_{i=1}^k (i-\frac{1}{2})\Gamma(\frac{1}{2})} \\
=& \frac{k!}{\pi^{\frac{1}{2}} \Pi_{i=1}^k (i-\frac{1}{2})} = \frac{(2k)!!}{\pi^{\frac{1}{2}}(2k-1)!!}
\end{aligned}$$

Recall the Wallis integral lemma that when d is even

$$\begin{aligned}
&\int_0^{\frac{\pi}{2}} \sin^d(t) dt = \int_0^{\frac{\pi}{2}} \cos^d(t) dt \\
=& \frac{\pi}{2} \cdot \frac{(d-1)!!}{d!!}, \quad d = 2k \in \mathbb{N}
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} =& \frac{(2k)!!}{\pi^{\frac{1}{2}}(2k-1)!!} = \frac{1}{2\pi^{-\frac{1}{2}} \cdot (\frac{\pi}{2} \cdot \frac{(2k-1)!!}{(2k)!!})} \\
=& \frac{1}{2\pi^{-\frac{1}{2}} \int_0^{\frac{\pi}{2}} \sin^{2k}(t) dt} = \frac{1}{2\pi^{-\frac{1}{2}} \int_0^{\frac{\pi}{2}} \sin^d(t) dt}
\end{aligned}$$

When d is odd, assume $d = 2k+1, k \in \mathbb{N}$ and note that $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$, then

$$\begin{aligned}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} =& \frac{\Gamma(k+\frac{3}{2})}{\Gamma(k+1)} = \frac{\Pi_{i=0}^{k}(i+\frac{1}{2})\Gamma(\frac{1}{2})}{k!\Gamma(1)} \\ =& \frac{\pi^{\frac{1}{2}}\Pi_{i=0}^{k}(i+\frac{1}{2})}{k!} = \frac{\pi^{\frac{1}{2}}(2k+1)!!}{2\cdot(2k)!!}\end{aligned}$$

Recall the Wallis integral lemma that when d is odd

$$\int_0^{\frac{\pi}{2}} \sin^d(t)dt = \int_0^{\frac{\pi}{2}} \cos^d(t)dt = \frac{(d-1)!!}{d!!}, \quad d = 2k+1 \in \mathbb{N}$$

Thus,

$$\begin{aligned}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} =& \frac{\pi^{\frac{1}{2}}(2k+1)!!}{2\cdot(2k)!!} = \frac{1}{2\pi^{-\frac{1}{2}}\cdot(\frac{(2k)!!}{(2k+1)!!})} \\ =& \frac{1}{2\pi^{-\frac{1}{2}}\int_0^{\frac{\pi}{2}}\sin^{2k+1}(t)dt} = \frac{1}{2\pi^{-\frac{1}{2}}\int_0^{\frac{\pi}{2}}\sin^{d}(t)dt}\end{aligned}$$

To sum up, for all $d \in \mathbb{N}$, we have

$$\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d+1}{2})} = \frac{1}{2\pi^{-\frac{1}{2}}\int_0^{\frac{\pi}{2}}\sin^{d}(t)dt} \tag{43}$$

Then we avoid the computation of $\Gamma(\frac{d}{2}+1), \Gamma(\frac{d+1}{2})$ and transfer it into the computation of an integral. Applying formula 43, we have

$$\begin{aligned}&\sup_{\|z\|_q \le \epsilon} TV(\mu, z+\mu) \\ =& 1 - \frac{1}{\int_0^{\frac{\pi}{2}}\sin^d(t)dt}\int_0^{\arccos(\frac{\max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\}}{2r})} \sin^d(t)dt\end{aligned}$$

□

I. Proof of Theorem 4.5

Proof. Recall the definition of l_p-norm constraint set of probability measures

$$\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \leq \epsilon\}$$

Note that

$$\begin{aligned}\sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu, \nu) &= \sup_{\|x-x'\|_q \leq \epsilon} TV(x + \mu, x' + \mu) \\ &= \sup_{\|z\|_q \leq \epsilon} TV(\mu, z + \mu)\end{aligned}$$

where the first equality is due to the definition of $\mathcal{D}_{x,\epsilon,q}$ and the second equality is due to the translation invariance property of total variation distance.

When $\epsilon \geq 2r$,

$$1 \geq \sup_{\|z\|_q \leq \epsilon} TV(\mu, z + \mu) \geq TV(\mu, \epsilon e_1 + \mu) = 1$$

where the first inequality is due to the fact that μ and $z + \mu$ are probability measures; the second inequality is due to supp$(\mu) \cap$ supp$(\epsilon e_1 + \mu) = \emptyset$. Thus, in this case,

$$\sup_{\|z\|_q \leq \epsilon} TV(\mu, z + \mu) = 1$$

When $\epsilon < 2r$,

$$\begin{aligned}&\sup_{\|z\|_q \leq \epsilon} TV(\mu, z + \mu) \\ &= \sup_{\|z\|_q \leq \epsilon} \frac{1}{2} \int \Big| \frac{1}{\text{Vol}(K)} \mathcal{I}_{x \in K} - \frac{1}{2\text{Vol}(K)} \mathcal{I}_{x \in z+K} \Big| dx \\ &= \sup_{\|z\|_q \leq \epsilon} \frac{1}{2\text{Vol}(K)} \int \mathcal{I}_{x \in K \Delta (z+K)} dx \\ &= \sup_{\|z\|_q \leq \epsilon} \frac{\text{Vol}(K \Delta (z + K))}{2\text{Vol}(K)} \\ &= \sup_{\|z\|_q \leq \epsilon} 1 - \frac{\text{Vol}(K \cap (z + K))}{\text{Vol}(K)} \\ &= \sup_{\|z\|_q \leq \epsilon} 1 - \frac{\Pi_{i=1}^d (2r - |z_i|)}{(2r)^d} \\ &= \sup_{\|z\|_q \leq \epsilon} 1 - \Pi_{i=1}^d \Big(1 - \frac{|z_i|}{2r}\Big)\end{aligned}$$

First, we study typical cases when $q = 1, 2, \infty$. When $q = 1$, we need to solve the following optimization problem

$$\inf_{\|z\|_1 \leq \epsilon} \Pi_{i=1}^d (2r - |z_i|)$$

Here we use mathematical induction to prove that

$$\inf_{\|z\|_1 \leq \epsilon} \Pi_{i=1}^d (2r - |z_i|) = (2r)^{d-1}(2r - \epsilon)$$

When $d = 2$,

$$\begin{aligned}\inf_{\|z\|_1 \leq \epsilon} \Pi_{i=1}^d (2r - |z_i|) &= \inf_{|z_1| + |z_2| \leq \epsilon} (2r - |z_1|)(2r - |z_2|) \\ &= \inf_{|z_2| \leq \epsilon} (2r - \epsilon + |z_2|)(2r - |z_2|) \\ &= \inf_{0 \leq z_2 \leq \epsilon} (2r - \epsilon + z_2)(2r - z_2) \\ &= \inf_{0 \leq z_2 \leq \epsilon} z_2(\epsilon - z_2) + 2r(2r - \epsilon) \\ &= 2r(2r - \epsilon)\end{aligned}$$

Thus, induction hypothesis holds for $d = 2$. Then, assume induction hypothesis holds for $d = n$. When $d = n + 1$,

$$\begin{aligned}
&\inf_{\|z\|_1 \le \epsilon} \Pi_{i=1}^{n+1}(2r - |z_i|) \\
=& \inf_{\sum_{i=1}^{n+1}|z_i| \le \epsilon} \Pi_{i=1}^{n+1}(2r - |z_i|) \\
=& \inf_{\sum_{i=1}^{n}|z_i| \le \epsilon - |z_{n+1}|} \big(\Pi_{i=1}^{n}(2r - |z_i|)\big)(2r - |z_{n+1}|) \\
=& \inf_{|z_{n+1}| \le \epsilon} (2r)^{n-1}(2r - \epsilon + |z_{n+1}|)(2r - |z_{n+1}|) && (44) \\
=& (2r)^n(2r - \epsilon) = (2r)^{d-1}(2r - \epsilon) && (45)
\end{aligned}$$

where 44 is due to the induction hypothesis when $d = n$; 45 is due to the induction hypothesis when $d = 2$. Therefore, we have already proved that

$$\inf_{\|z\|_1 \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) = (2r)^{d-1}(2r - \epsilon), \forall d \in \mathbb{N}$$

Plugging in this result, it follows that

$$\begin{aligned}
\sup_{\|z\|_1 \le \epsilon} TV(\mu, z + \mu) = & \sup_{\|z\|_1 \le \epsilon} 1 - \frac{\Pi_{i=1}^{d}(2r - |z_i|)}{(2r)^d} \\
=& 1 - \frac{(2r)^{d-1}(2r - \epsilon)}{(2r)^d} = \frac{\epsilon}{2r}
\end{aligned}$$

When $q = 2$, we need to solve the following optimization problem

$$\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) \tag{46}$$

When $d = 2$,

$$\begin{aligned}
\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) =& \inf_{|z_1|^2 + |z_2|^2 \le \epsilon^2} (2r - |z_1|)(2r - |z_2|) \\
=& \inf_{|z_2| \le \epsilon} \left(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\right)(2r - |z_2|) \\
=& \inf_{0 \le z_2 \le \epsilon} \left(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\right)(2r - z_2)
\end{aligned}$$

Define $f(z_2) = \ln\left(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\right) + \ln(2r - z_2)$, then

$$\begin{aligned}
f'(z_2) =& \frac{z_2(\epsilon^2 - z_2^2)^{-\frac{1}{2}}}{2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}} - \frac{1}{2r - z_2} \\
=& \frac{2rz_2(\epsilon^2 - z_2^2)^{-\frac{1}{2}} - z_2^2(\epsilon^2 - z_2^2)^{-\frac{1}{2}} - 2r + (\epsilon^2 - z_2^2)^{\frac{1}{2}}}{\left(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\right)(2r - z_2)}
\end{aligned}$$

Define $g(z_2) = 2rz_2(\epsilon^2 - z_2^2)^{-\frac{1}{2}} - z_2^2(\epsilon^2 - z_2^2)^{-\frac{1}{2}} - 2r + (\epsilon^2 - z_2^2)^{\frac{1}{2}}$, then

$$g'(z_2) = (2z_2^3 - 3\epsilon^2 z_2 + 2r\epsilon^2)(\epsilon^2 - z_2^2)^{-\frac{3}{2}}$$

Define $h(z_2) = 2z_2^3 - 3\epsilon^2 z_2 + 2r\epsilon^2$, then $h'(z_2) = 6z_2^2 - 3\epsilon^2 = 6(z_2 - \frac{\epsilon}{\sqrt{2}})(z_2 + \frac{\epsilon}{\sqrt{2}})$. Thus, when $0 \le z_2 \le \frac{\epsilon}{\sqrt{2}}$, $h'(x) \le 0$; when $\frac{\epsilon}{\sqrt{2}} < z_2 \le \epsilon$, $h'(x) > 0$. Thus, the minimum value of $h(x)$ on interval $[0, \epsilon]$ is $h(\frac{\epsilon}{\sqrt{2}}) = \sqrt{2}\epsilon^2(\sqrt{2}r - \epsilon)$. Therefore, function $f(z_2)$ behaves differently when $0 < \epsilon \le \sqrt{2}r$ and when $\sqrt{2}r < \epsilon < 2r$.

When $0 < \epsilon \le \sqrt{2}r$, $h(z_2) \ge h(\frac{\epsilon}{\sqrt{2}}) = \sqrt{2}\epsilon^2(\sqrt{2}r - \epsilon) \ge 0$ on interval $[0, \epsilon]$ and therefore $g'(z_2) = h(z_2)(\epsilon^2 - z_2^2)^{-\frac{3}{2}} \ge 0$. Note that $g(0) = \epsilon - 2r < 0, g(\frac{\epsilon}{\sqrt{2}}) = 0, g(\epsilon^-) = \infty$ and therefore $f'(z_2) \le 0$ when $0 \le z_2 \le \frac{\epsilon}{\sqrt{2}}$ while $f'(z_2) > 0$ when $\frac{\epsilon}{\sqrt{2}} < z_2 \le \epsilon$. Thus, $f(z_2)$ takes its minimum when $z_2 = \frac{\epsilon}{\sqrt{2}}$. In this case,

$$\inf_{0 \le z_2 \le \epsilon} \left(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\right)(2r - z_2) = (2r - \frac{\epsilon}{\sqrt{2}})^2 \tag{47}$$

When $\sqrt{2}r < \epsilon < 2r$, we have $h(0) = 2r\epsilon^2 > 0, h(\frac{\epsilon}{\sqrt{2}}) = \sqrt{2}\epsilon^2(\sqrt{2}r - \epsilon) < 0, h(\epsilon) = \epsilon^2(2r - \epsilon) > 0$. Assume $h(t_1) = h(t_2) = 0, 0 < t_1 < \frac{\epsilon}{\sqrt{2}} < t_2 < \epsilon$, then when $0 \le z_2 \le t_1$ or $t_2 \le z_2 \le \epsilon$, $h(z_2) \ge 0$ and when $t_1 < z_2 < t_2$, $h(z_2) < 0$. Therefore, $g'(z_2) \ge 0$ when $0 \le z_2 \le t_1$ or $t_2 \le z_2 \le \epsilon$; $g'(z_2) < 0$ when $t_1 < z_2 < t_2$. Note that $g(z_2) = 0 \iff (2z_2^2 - \epsilon^2)\big(2(z_2 - r)^2 + 2r^2 - \epsilon^2\big) = 0$, therefore when $0 \le z_2 \le r - \sqrt{\frac{\epsilon^2}{2} - r^2}$ or $\frac{\epsilon}{\sqrt{2}} \le z_2 \le r + \sqrt{\frac{\epsilon^2}{2} - r^2}$, $g(z_2) \le 0$; when $r - \sqrt{\frac{\epsilon^2}{2} - r^2} < z_2 < \frac{\epsilon}{\sqrt{2}}$ or $r + \sqrt{\frac{\epsilon^2}{2} - r^2} < z_2 < \epsilon$, $g(z_2) > 0$. Thus, when $0 \le z_2 \le r - \sqrt{\frac{\epsilon^2}{2} - r^2}$ or $\frac{\epsilon}{\sqrt{2}} \le z_2 \le r + \sqrt{\frac{\epsilon^2}{2} - r^2}$, $f'(x) \le 0$; when $r - \sqrt{\frac{\epsilon^2}{2} - r^2} < z_2 < \frac{\epsilon}{\sqrt{2}}$ or $r + \sqrt{\frac{\epsilon^2}{2} - r^2} < z_2 < \epsilon$, $f'(x) > 0$. Thus, $f(z_2)$ takes its minimum when $z_2 = r - \sqrt{\frac{\epsilon^2}{2} - r^2}$ or $z_2 = r + \sqrt{\frac{\epsilon^2}{2} - r^2}$. In this case,

$$\begin{aligned} &\inf_{0 \le z_2 \le \epsilon} \Big(2r - (\epsilon^2 - z_2^2)^{\frac{1}{2}}\Big)(2r - z_2) \\ =&\Big(r - \sqrt{\frac{\epsilon^2}{2} - r^2}\Big)\Big(r + \sqrt{\frac{\epsilon^2}{2} - r^2}\Big) = 2r^2 - \frac{\epsilon^2}{2} \\ &\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) = \inf_{\sum_{i=1}^{n+1} z_i^2 \le \epsilon^2} \Pi_{i=1}^{n+1}(2r - |z_i|) \\ =& \inf_{\sum_{i=1}^{n} z_i^2 \le \epsilon^2 - z_{n+1}^2} \Big(\Pi_{i=1}^{n}(2r - |z_i|)\Big)(2r - |z_{n+1}|) \end{aligned} \tag{48}$$

By then, we have understand clearly the optimization problem when $d = 2$.

Then, consider the case when $d = 3$. When $d = 3$,

$$\begin{aligned} &\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) \\ =& \inf_{z_1^2 + z_2^2 + z_3^2 \le \epsilon^2} (2r - |z_1|)(2r - |z_2|)(2r - |z_3|) \end{aligned} \tag{49}$$

When $0 < \epsilon \le \sqrt{2}r$, assume the optimal point is z^*. We will prove that each coordinate of z^* has the same value. Here we use reduction to absurdity, and wlog assume $z_1^* \ne z_2^*$. By fixing the value of z_3^*, the optimization problem 46 is equivalent to

$$\inf_{z_1^2 + z_2^2 \le \epsilon^2 - (z_3^*)^2} (2r - |z_1|)(2r - |z_2|)$$

And (z_1^*, z_2^*) should be an optimal point of above problem. Note that $\epsilon^2 - (z_3^*)^2 \le \epsilon^2 \le 2r^2$ and applying 47, we know that $z_1^* = z_2^*$ which is a contradiction. Thus, $z_1^* = z_2^* = z_3^* = c$. And

$$\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{n}(2r - |z_i|) = \inf_{c \le \frac{\epsilon}{\sqrt{3}}} (2r - c)^3 = \Big(2r - \frac{\epsilon}{\sqrt{3}}\Big)^3$$

When $\sqrt{2}r < \epsilon \le \sqrt{3}r$, it's obvious that the optimal point z^* of optimization problem 49 must lie on the boundary of feasible region, i.e. $(z_1^*)^3 + (z_2^*)^3 + (z_3^*)^3 = \epsilon^2$. Wlog, assume $(z_3^*)^3 \ge \frac{\epsilon^2}{3}$ and $(z_1^*)^2 + (z_2^*)^2 \le \frac{2\epsilon^2}{3} \le 2r^3$. By fixing the value of z_3^* and following similar deduction procedure as above we know that $z_1^* = z_2^* = c^*$, where c^* is the optimal point of following optimization problem.

$$\inf_{0 \le c \le \frac{\epsilon}{\sqrt{3}}} (2r - c)^2(2r - \sqrt{\epsilon^2 - 2c^2}) \tag{50}$$

Define $f(x) = 2\ln(2r - x) + \ln(2r - \sqrt{\epsilon^2 - 2x^2})$ where $0 \le x \le \frac{\epsilon}{\sqrt{3}}$, then

$$f'(x) = \frac{2(3x^2 - 2rx - \epsilon^2 + 2r\sqrt{\epsilon^2 - 2x^2})}{(x - 2r)(2r - \sqrt{\epsilon^2 - 2x^2})\sqrt{\epsilon^2 - 2x^2}}$$

It's obvious that the denominator of $f'(x)$ is negative. As for the numerator, define $g(x) = 3x^2 - 2rx - \epsilon^2$ where $0 \le x \le \frac{\epsilon}{\sqrt{3}}$. Note that

$$g(x) \le \max\Big\{g(0), g\Big(\frac{\epsilon}{\sqrt{3}}\Big)\Big\} = \max\Big\{-\epsilon^2, -\frac{2r\epsilon}{\sqrt{3}}\Big\} \le 0$$

Thus, we have the following equivalent relationship

$$
\begin{aligned}
& 3x^2 - 2rx - \epsilon^2 + 2r\sqrt{\epsilon^2 - 2x^2} \leq 0 \\
\Longleftrightarrow\ & 3x^2 - 2rx - \epsilon^2 \leq -2r\sqrt{\epsilon^2 - 2x^2} \leq 0 \\
\Longleftrightarrow\ & (3x^2 - 2rx - \epsilon^2)^2 \geq \big(-2r\sqrt{\epsilon^2 - 2x^2}\big)^2 \geq 0 \\
\Longleftrightarrow\ & (3x^2 - \epsilon^2)(3x^2 - 4rx + 4r^2 - \epsilon^2) \geq 0 \\
\Longleftrightarrow\ & 3x^2 - 4rx + 4r^2 - \epsilon^2 \leq 0 \\
\Longleftrightarrow\ & \begin{cases} \emptyset, \sqrt{2}r < \epsilon \leq 2\sqrt{\dfrac{2}{3}}r \\ \dfrac{2r - \sqrt{3\epsilon^2 - 8r^2}}{3} \leq x \leq \dfrac{2r + \sqrt{3\epsilon^2 - 8r^2}}{3}, 2\sqrt{\dfrac{2}{3}}r < \epsilon \leq \sqrt{3}r \end{cases}
\end{aligned}
$$

where the last equivalent relationship is due to the discriminant of the quadratic equation $3x^2-4rx+4r^2-\epsilon^2$ is $\Delta = 4(3\epsilon^2 - 8r^2)$. Therefore, when $\sqrt{2}r < \epsilon \leq 2\sqrt{\frac{2}{3}}r$, $f'(x) \leq 0, \forall 0 \leq x \leq \frac{\epsilon}{\sqrt{3}}$ and hence the optimal point c^* in the optimization problem 50 takes value $\frac{\epsilon}{\sqrt{3}}$, whereas when $2\sqrt{\frac{2}{3}}r < \epsilon \leq \sqrt{3}r$, $f'(x) \leq 0$ for $0 \leq x \leq \frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$, $\frac{2r+\sqrt{3\epsilon^2-8r^2}}{3} \leq x \leq \frac{\epsilon}{\sqrt{3}}$ and $f'(x) > 0$ for $\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3} \leq x \leq \frac{2r+\sqrt{3\epsilon^2-8r^2}}{3}$ and note that $f(\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}) < f(\frac{\epsilon}{\sqrt{3}})$ hence the optimal point c^* in the optimization problem takes value $\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$. To sum up, when $\sqrt{2}r < \epsilon \leq 2\sqrt{\frac{2}{3}}r$, the optimal point z^* of optimization problem 49 satisfies $z_1^* = z_2^* = z_3^* = \frac{\epsilon}{\sqrt{3}}$. And when $2\sqrt{\frac{2}{3}}r < \epsilon \leq \sqrt{3}r$, the optimal point z^* of optimization problem 49 satisfies $z_1^* = z_2^* = \frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$, $z_3^* = \frac{4r+\sqrt{3\epsilon^2-8r^2}}{3}$ or one of its permutations.

When $\sqrt{3}r < \epsilon < 2r$, similarly we have $(z_1^*)^3 + (z_2^*)^3 + (z_3^*)^3 = \epsilon^2$. If there exists $1 \leq i \leq 3$ such that $(z_i^*)^2 \geq \epsilon^2 - 2r^2$, wlog assume $(z_3^*)^2 \geq \epsilon^2 - 2r^2$. By substituting the value range of x from $[0, \frac{\epsilon}{\sqrt{3}}]$ into $[0, r]$, following similar deduction procedure and noticing that $f(\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}) < f(r)$, we know that the optimal point z^* in this case satisfies $z_1^* = z_2^* = \frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$, $z_3^* = \frac{4r+\sqrt{3\epsilon^2-8r^2}}{3}$ or one of its permutations. On the other hand, if $(z_i^*)^2 < \epsilon^2 - 2r^2$ for all $1 \leq i \leq 3$, then $(z_1^*)^2 + (z_2^*)^2 = \epsilon^2 - (z_3^*)^2 > 2r^2$. Applying 48 and taking $z_1^* = r - \sqrt{\frac{\epsilon^2-(z_3^*)^2}{2} - r^2}$, $z_2^* = r + \sqrt{\frac{\epsilon^2-(z_3^*)^2}{2} - r^2}$, we know the optimization problem is equivalent to

$$
\inf_{0 \leq z_3 < \sqrt{\epsilon^2 - 2r^2}} \Big(2r^2 - \frac{\epsilon^2 - z_3^2}{2}\Big)(2r - z_3)
$$

According to monotonicity analysis of the cubic function above, the optimal point z_3^* is either $\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$ or $\sqrt{\epsilon^2 - 2r^2}$. And it's easy to verify that $f(\frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}) < f(\sqrt{\epsilon^2 - 2r^2})$ and therefore $z_3^* = \frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$. However, $(z_2^*)^2 > \epsilon^2 - 2r^2$ which leads to a contradiction.

In summary, considering the case $d = 3$, when $0 < \epsilon \leq 2\sqrt{\frac{2}{3}}r$, the optimal point z^* of original optimization problem satisfies $z_1^* = z_2^* = z_3^* = \frac{\epsilon}{\sqrt{3}}$ and the optimal value is $(2r - \frac{\epsilon}{\sqrt{3}})^3$ and when $\sqrt{3}r < \epsilon < 2r$, the optimal point z^*, the optimal point z^* of original optimization problem satisfies $z_1^* = z_2^* = \frac{2r-\sqrt{3\epsilon^2-8r^2}}{3}$, $z_3^* = \frac{4r+\sqrt{3\epsilon^2-8r^2}}{3}$ or one of its permutations.

Next, consider the general case when $d = n \geq 4$. In the first place, we point out and prove two useful properties of the optimal point z^* which help simplify our later discussion a lot.

- All coordinates of optimal point z^* takes at most two different values.
- If the coordinates of an optimal point z^* takes exactly two different values c_1 and c_2, then the number of coordinates equal to c_1 must be $n-1$ or 1.

Proof. On one hand, by using reduction to absurdity, wlog assume z_1^*, z_2^*, z_3^* take three different values. Fixing the value of the other $n-3$ coordinates, we know that (z_1^*, z_2^*, z_3^*) is the optimal point of a special case of original problem when $d = 3$.

And note that for all the optimal points of $d = 3$, there must exist two coordinates taking the same value, which leads to a contradiction. Thus, the first property is satisfied.
On the other hand, similarly, by applying reduction to absurdity, wlog assume $z_1^* = z_2^* = c_1$ and $z_3^* = z_4^* = c_2$ where $c_1 \neq c_2$. Fixing z_2^*, z_4^* and the value of the other $n-4$ coordinates and aware of the fact that $(z_1^*)^2 + (z_3^*)^2 \leq \frac{\epsilon^2}{2} < 2r^2$, we know that (z_1^*, z_3^*) is the optimal point of a special case of original problem when $d = 2, \epsilon < \sqrt{2}r$ and therefore $z_1^* = z_3^*$, which leads to a contradiction. Thus, the second property is satisfied. □

Using the two properties above, we know that the optimal point z^* has only two possible forms: $z^* = \left(\frac{\epsilon}{\sqrt{n}}, \cdots, \frac{\epsilon}{\sqrt{n}}\right)$ and $z^* = \left(c, \cdots, c, \sqrt{\epsilon^2 - (n-1)c^2}\right)$ or one of its permutations where $0 \leq c \leq \frac{\epsilon}{\sqrt{n-1}}, c \neq \frac{\epsilon}{\sqrt{n}}$, which can be unified into one form: $z^* = \left(c, \cdots, c, \sqrt{\epsilon^2 - (n-1)c^2}\right)$ or one of its permutations where $0 \leq c \leq \frac{\epsilon}{\sqrt{n-1}}$. Thus, the original problem can be simplified into following optimization problem with one degree of freedom:

$$\inf_{0 \leq c \leq \frac{\epsilon}{\sqrt{n-1}}} (2r - c)^{n-1}\left(2r - \sqrt{\epsilon^2 - (n-1)c^2}\right)$$

Define $f(x) = (n-1)\ln(2r - x) + \ln\left(2r - \sqrt{\epsilon^2 - (n-1)x^2}\right)$ where $0 \leq x \leq \frac{\epsilon}{\sqrt{n-1}}$, then

$$\begin{aligned} f'(x) &= \frac{(n-1)\left(nx^2 - 2rx - \epsilon^2 + 2r\sqrt{\epsilon^2 - (n-1)x^2}\right)}{(x - 2r)\left(2r - \sqrt{\epsilon^2 - (n-1)x^2}\right)\sqrt{\epsilon^2 - (n-1)x^2}}, \\ &\text{where } 0 \leq x < \frac{\epsilon}{\sqrt{n-1}} \end{aligned}$$

It's obvious that the denominator of $f'(x)$ is negative. As for the numerator, define $g(x) = nx^2 - 2rx - \epsilon^2$ where $0 \leq x \leq \frac{\epsilon}{\sqrt{n-1}}$. Note that

$$\begin{aligned} g(x) &\leq \min\left\{g(0), g\Big(\frac{\epsilon}{\sqrt{n-1}}\Big)\right\} \\ &= \max\left\{0, \frac{\epsilon(\epsilon - 2\sqrt{n-1}r)}{n-1}\right\} \leq 0 \end{aligned}$$

where the last inequality is due to $\epsilon < 2r < 2\sqrt{n-1}r$. Thus, when $0 \leq x \leq \frac{\epsilon}{\sqrt{n-1}}$,

$$\begin{aligned} & nx^2 - 2rx - \epsilon^2 + 2r\sqrt{\epsilon^2 - (n-1)x^2} \leq 0 \\ \Longleftrightarrow\ & nx^2 - 2rx - \epsilon^2 \leq -2r\sqrt{\epsilon^2 - (n-1)x^2} \leq 0 \\ \Longleftrightarrow\ & (nx^2 - 2rx - \epsilon^2)^2 \geq \left(-2r\sqrt{\epsilon^2 - (n-1)x^2}\right)^2 \\ \Longleftrightarrow\ & (nx^2 - \epsilon^2)(nx^2 - 4rx + 4r^2 - \epsilon^2) \geq 0 \\ \Longleftrightarrow\ & \Big(x - \frac{\epsilon}{\sqrt{n}}\Big)(nx^2 - 4rx + 4r^2 - \epsilon^2) \geq 0 \\ \Longleftrightarrow\ & \begin{cases} x \geq \dfrac{\epsilon}{\sqrt{n}} \text{ when } 0 < \epsilon < 2\sqrt{\dfrac{n-1}{n}}r \\ x \geq \dfrac{\epsilon}{\sqrt{n}}, \dfrac{2r - \sqrt{n\epsilon^2 - 4(n-1)r^2}}{n} \leq x \\ \leq \dfrac{2r + \sqrt{n\epsilon^2 - 4(n-1)r^2}}{n} \text{ when } 2\sqrt{\dfrac{n-1}{n}}r \leq \epsilon < 2r \end{cases} \end{aligned}$$

where the last equivalence relationship is due to the discriminant of the quadratic equation $nx^2 - 4rx + 4r^2 - \epsilon^2 = 0$ is

$$\Delta = 4\Big(n\epsilon^2 - 4(n-1)r^2\Big) < 0 \iff 0 < \epsilon < 2\sqrt{\frac{n-1}{n}}r$$

Thus, if $0 < \epsilon < 2\sqrt{\frac{n-1}{n}}r$, then $f'(x) \geq 0$ when $\frac{\epsilon}{\sqrt{n}} \leq x \leq \frac{\epsilon}{\sqrt{n-1}}$ and $f'(x) < 0$ when $0 \leq x < \frac{\epsilon}{\sqrt{n}}$. Thus, $f(x)$ takes its minimum when $x = \frac{\epsilon}{\sqrt{n}}$ and therefore $c^* = \frac{\epsilon}{\sqrt{n}}$.

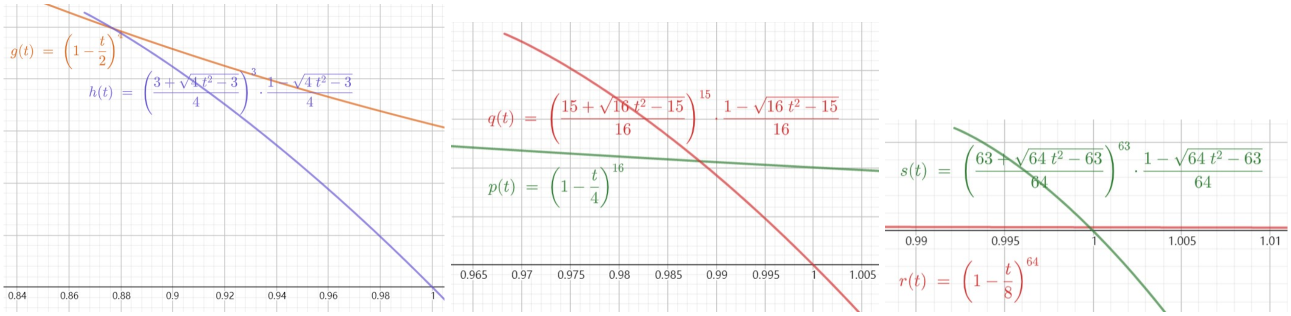

Figure 6. Graphs of functions $f_1(t) = \left(1 - \frac{t}{\sqrt{n}}\right)^n$, $f_2(t) = \left(\frac{(n-1)+\sqrt{nt^2-(n-1)}}{n}\right)^{n-1}\left(\frac{1-\sqrt{nt^2-(n-1)}}{n}\right)$ when $n = 4, 16, 64$ from left to right. According to the figure, on interval $\left[\sqrt{\frac{n-1}{n}}, 1\right]$, $f_2(t)$ is greater than $f_1(t)$ at first and then $f_2(t)$ exceeds $f_1(t)$. Furthermore, as n increases, the horizontal coordinate of the intersection point converge to 1, which can be seen intuitively from the figure above.

If $2\sqrt{\frac{n-1}{n}}r \leq \epsilon < 2r$, then $f'(x) \geq 0$ when $\frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n} \leq x \leq \frac{2r+\sqrt{n\epsilon^2-4(n-1)r^2}}{n}$ or $\frac{\epsilon}{\sqrt{n}} \leq x \leq \frac{\epsilon}{\sqrt{n-1}}$ and $f'(x) < 0$ when $0 \leq x < \frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n}$ or $\frac{2r+\sqrt{n\epsilon^2-4(n-1)r^2}}{n} < x < \frac{\epsilon}{\sqrt{n}}$. In this case, $f(x)$ takes its minimum when $x = \frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n}$ or $x = \frac{\epsilon}{\sqrt{n}}$. For the convenience of analysis, assume $t = \frac{\epsilon}{2r}$, $\sqrt{\frac{n-1}{n}} \leq t < 1$ and it follows that

$$e^{f\left(\frac{\epsilon}{\sqrt{n}}\right)} = (2r)^n\left(1 - \frac{t}{\sqrt{n}}\right)^n$$

$$\begin{aligned}&e^{f\left(\frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n}\right)}\\=&(2r)^n\left(\frac{(n-1)+\sqrt{nt^2-(n-1)}}{n}\right)^{n-1}\cdot\left(\frac{1-\sqrt{nt^2-(n-1)}}{n}\right)\end{aligned}$$

We can prove that there exists $t_n \in \left[\sqrt{\frac{n-1}{n}}, 1\right)$ such that $c^* = \frac{\epsilon}{\sqrt{n}}$ when $2\sqrt{\frac{n-1}{n}}r \leq \epsilon \leq 2t_n r$ and $c^* = \frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n}$ when $2t_n r < \epsilon < 2r$ while t_n converge to 1 at an exponential rate as shown in figure 6.

In conclusion, for the case $d = n \geq 4$, when $0 < \epsilon \leq 2t_n r$, $c^* = \frac{\epsilon}{\sqrt{n}}$ and therefore

$$\begin{aligned}&\inf_{\|z\|_2\leq\epsilon} \Pi_{i=1}^n(2r-|z_i|)\\=&(2r-c^*)^{n-1}\left(2r-\sqrt{\epsilon^2-(n-1)(c^*)^2}\right)\\=&\left(2r-\frac{\epsilon}{n^{\frac{1}{2}}}\right)^n = \left(2r-\frac{\epsilon}{d^{\frac{1}{2}}}\right)^d\end{aligned}$$

Plugging in this result, it follows that

$$\begin{aligned}&\sup_{\|z\|_2\leq\epsilon} TV(\mu, z+\mu)\\=&\sup_{\|z\|_2\leq\epsilon} 1-\frac{\Pi_{i=1}^d(2r-|z_i|)}{(2r)^d}\\=&1-\frac{\left(2r-\frac{\epsilon}{d^{\frac{1}{2}}}\right)^d}{(2r)^d}\\=&1-\left(1-\frac{\epsilon}{2d^{\frac{1}{2}}r}\right)^d\end{aligned}$$

and when $2t_n r < \epsilon < 2r$, $c^* = \frac{2r-\sqrt{n\epsilon^2-4(n-1)r^2}}{n}$ and therefore

$$
\begin{aligned}
&\inf_{\|z\|_2 \le \epsilon} \Pi_{i=1}^{n}(2r - |z_i|) \\
&= (2r - c^*)^{n-1}\big(2r - \sqrt{\epsilon^2 - (n-1)(c^*)^2}\big) \\
&= \left(\frac{2(n-1)r + \sqrt{n\epsilon^2 - 4(n-1)r^2}}{n}\right)^{n-1}\left(\frac{2r - \sqrt{n\epsilon^2 - 4(n-1)r^2}}{n}\right) \\
&= \left(\frac{2(d-1)r + \sqrt{d\epsilon^2 - 4(d-1)r^2}}{n}\right)^{n-1}\left(\frac{2r - \sqrt{d\epsilon^2 - 4(d-1)r^2}}{d}\right)
\end{aligned}
$$

Plugging in this result, it follows that

$$
\begin{aligned}
&\sup_{\|z\|_2 \le \epsilon} TV(\mu, z + \mu) \\
&= \sup_{\|z\|_2 \le \epsilon} 1 - \frac{\Pi_{i=1}^{d}(2r - |z_i|)}{(2r)^d} \\
&= 1 - \frac{\left(\frac{2(d-1)r+\sqrt{d\epsilon^2-4(d-1)r^2}}{d}\right)^{d-1}\left(\frac{2r-\sqrt{d\epsilon^2-4(d-1)r^2}}{d}\right)}{(2r)^d} \\
&= 1 - \left(\frac{d - 1 + \sqrt{d(\frac{\epsilon}{2r})^2 - d + 1}}{d}\right)^{d-1}\left(\frac{1 - \sqrt{d(\frac{\epsilon}{2r})^2 - d + 1}}{d}\right)
\end{aligned}
$$

When $q = \infty$, it's easy to verify that

$$
\inf_{\|z\|_\infty \le \epsilon} \Pi_{i=1}^{d}(2r - |z_i|) = (2r - \epsilon)^d
$$

Plugging in the formula above, it follows

$$
\begin{aligned}
\sup_{\|z\|_\infty \le \epsilon} TV(\mu, z + \mu) = &\sup_{\|z\|_\infty \le \epsilon} 1 - \frac{\Pi_{i=1}^{d}(2r - |z_i|)}{(2r)^d} \\
=&1 - \frac{(2r - \epsilon)^d}{(2r)^d} = 1 - \left(1 - \frac{\epsilon}{2r}\right)^d
\end{aligned}
$$

□

J. Proof of Theorem 4.6

Proof. Recall the definition of l_p-norm constraint set of probability measures

$$
\mathcal{D}_{x,\epsilon,q} = \{x' + \mu : \|x - x'\|_q \le \epsilon\}
$$

Assume $\mathcal{D}_{x,\epsilon,q} \subseteq \mathcal{D}_{x,\xi(\epsilon)}$, then

$$
\begin{aligned}
\xi(\epsilon) \ge \sup_{\nu \in \mathcal{D}_{x,\epsilon,q}} TV(\mu, \nu) &= \sup_{\|x - x'\|_q \le \epsilon} TV(x + \mu, x' + \mu) \\
= \sup_{\|z\|_q \le \epsilon} TV(\mu, z + \mu) &\ge TV(\mu, \epsilon e_1 + \mu)
\end{aligned}
$$

which indicates that $TV(\mu, \epsilon e_1 + \mu)$ provides a lower bound for $\xi(\epsilon)$. Thus, we only need to estimate the value of $TV(\mu, \epsilon e_1 + \mu)$. According to lemma B.3, we have

$$
TV(\mu, \epsilon e_1 + \mu) = \frac{\text{Vol}(K\Delta(\epsilon e_1 + K))}{2\text{Vol}(K)} = 1 - \frac{\text{Vol}(K \cap (\epsilon e_1 + K))}{\text{Vol}(K)}
$$

Note that

$$\begin{aligned}
&K \cap (\epsilon e_1 + K) \\
&= \left\{ x \in \mathbb{R}^d \middle| |x_1|^p + \cdots + |x_d|^p \leq r^p, |x_1 - \epsilon|^p + |x_2|^p + \cdots + |x_d|^p \leq r^p \right\} \\
&= \left\{ x \in \mathbb{R}^d \middle| \epsilon - (r^p - (|x_2|^p + \cdots + |x_d|^p|))^{\frac{1}{p}} \leq x_1 \leq (r^p - (|x_2|^p + \cdots + |x_d|^p|))^{\frac{1}{p}} \right\} \\
&= \left\{ x \in \mathbb{R}^d \middle| \epsilon - (r^p - (|x_2|^p + \cdots + |x_d|^p))^{\frac{1}{p}} \leq x_1 \leq \frac{\epsilon}{2} \right\} \cup \left\{ x \middle| \frac{\epsilon}{2} \leq x_1 \leq (r^p - (|x_2|^p + \cdots + |x_d|^p|))^{\frac{1}{p}} \right\} \\
&:= \Omega_1 \cup \Omega_2 \text{ where } \Omega_1 \cap \Omega_2 = \emptyset
\end{aligned}$$

It's easy to verify that $\text{Vol}(\Omega_1) = \text{Vol}(\Omega_2)$ according to integration by substitution and therefore $\text{Vol}(K \cap (\epsilon e_1 + K)) = 2\text{Vol}(\Omega_2)$. To estimate the volume of Ω_2, we first introduce several lemmas below for the convenience of later discussion.

Lemma J.1 (Volume formula of d-dimensional l_p norm ball)**.** *The volume of d-dimensional l_p ball of radius r is given by*

$$V_p^{(d)} = (2r)^d \frac{\Gamma(1 + \frac{1}{p})^d}{\Gamma(1 + \frac{d}{p})}$$

Lemma J.2. *The d-dimensional l_p ball of volume 1 has radius about $\frac{d^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}$.*

Proof. When dimension d is big enough, we can obtain an asymptotic volume estimation of l_p norm ball with radius r.

$$\begin{aligned}
V_p^{(d)} =& (2r)^d \frac{\Gamma(1 + \frac{1}{p})^d}{\Gamma(1 + \frac{d}{p})} \approx (2r)^d \frac{\Gamma(1 + \frac{1}{p})^d}{\sqrt{2\pi \frac{d}{p}} \left(\frac{d}{pe}\right)^{\frac{d}{p}}} \\
=& \sqrt{\frac{p}{2\pi d}} \left(\frac{2r(pe)^{\frac{1}{p}} \Gamma(1 + \frac{1}{p})}{d^{\frac{1}{p}}} \right)^d
\end{aligned}$$

where the first equality is due to lemma J.1 and the approximate equality is due to Stirling's formula about the estimation of gamma function that $\Gamma(z + 1) \approx \sqrt{2\pi z} \left(\frac{z}{e}\right)^z$. Thus, when $V_p^{(d)} = 1$, we have

$$r \approx \frac{d^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}} \Gamma(1 + \frac{1}{p})}$$

□

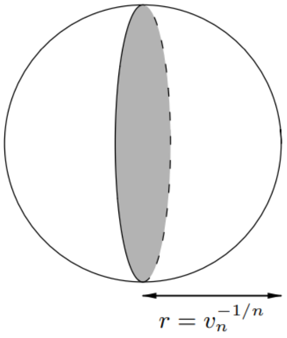

Figure 7. Comparing the volume of a ball with that of its central slice

Then we estimate the volume of l_p norm ball cap by studying the asymptotic property of the mass distribution of l_p norm ball. To begin with, let's estimate the $(d-1)$-dimensional volume of a slice through the center of the l_p ball of volume 1. Note that the ball has radius $r = (V_p^{(d)})^{-\frac{1}{d}}$. The slice is an $(d-1)$-dimensional ball of this radius, so its volume is

$$
\begin{aligned}
V_p^{(d-1)} r^{d-1} &= V_p^{(d-1)} (V_p^{(d)})^{-\frac{d-1}{d}} \\
&= 2^{d-1} \frac{\Gamma(1+\frac{1}{p})^{d-1}}{\Gamma(1+\frac{d-1}{p})} \left(2^d \frac{\Gamma(1+\frac{1}{p})^d}{\Gamma(1+\frac{d}{p})} \right)^{-\frac{d-1}{d}}
\end{aligned}
$$

Using Stirling's formula again, when d is sufficiently large, we have

$$
\begin{aligned}
& V_p^{(d-1)} r^{d-1} \\
&= 2^{d-1} \frac{\Gamma(1+\frac{1}{p})^{d-1}}{\Gamma(1+\frac{d-1}{p})} \left(2^d \frac{\Gamma(1+\frac{1}{p})^d}{\Gamma(1+\frac{d}{p})} \right)^{-\frac{d-1}{d}} \\
&= 2^{d-1} \frac{\Gamma(1+\frac{1}{p})^{d-1}}{\sqrt{2\pi\frac{d-1}{p}}\left(\frac{d-1}{pe}\right)^{\frac{d-1}{p}}} \left(2^d \frac{\Gamma(1+\frac{1}{p})^d}{\sqrt{2\pi\frac{d}{p}}\left(\frac{d}{pe}\right)^{\frac{d}{p}}} \right)^{-\frac{d-1}{d}} \\
&= \frac{1}{\sqrt{2\pi\frac{d-1}{p}}\left(\frac{d-1}{pe}\right)^{\frac{d-1}{p}}} \cdot \frac{1}{\left(\sqrt{2\pi\frac{d}{p}}\left(\frac{d}{pe}\right)^{\frac{d}{p}}\right)^{-\frac{d-1}{d}}} \\
&= \frac{1}{(\frac{2\pi}{p})^{\frac{1}{2d}} (d-1)^{\frac{d-1}{p}+\frac{1}{2}} d^{-\frac{d-1}{p}-\frac{d-1}{2d}}} \\
&= \frac{(1+\frac{1}{d-1})^{\frac{d-1}{p}+\frac{1}{2}}}{(\frac{2\pi d}{p})^{\frac{1}{2d}}} \approx e^{\frac{1}{p}}
\end{aligned}
$$

where the second equality is due to Stirling's formula for $\Gamma(1+\frac{d-1}{p})$ and $\Gamma(1+\frac{d}{p})$; the third equality just eliminate the exponential of 2 and $\Gamma(1+\frac{1}{p})$. Thus, we conclude that the slice has volume about $e^{\frac{1}{p}}$ when d is large.

Then, consider the $(d-1)$-dimensional volumes of parallel slices. The slice at distance x from the center is an $(d-1)$-dimensional ball whose radius is $(r^p - x^p)^{\frac{1}{p}}$, so the volume of the smaller slice is about

$$
e^{\frac{1}{p}} \left(\frac{(r^p - x^p)^{\frac{1}{p}}}{r} \right)^{d-1} = e^{\frac{1}{p}} \left(1 - \left(\frac{x}{r}\right)^p \right)^{\frac{d-1}{p}}
$$

Since r is roughly $\frac{d^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}$, this is about

$$\begin{aligned}
&e^{\frac{1}{p}}\left(1-\Big(\frac{2x(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}{d^{\frac{1}{p}}}\Big)^{p}\right)^{\frac{d-1}{p}}\\
=&e^{\frac{1}{p}}\left(1-\frac{pe(2x\Gamma(1+\frac{1}{p}))^{p}}{d}\right)^{\frac{d-1}{p}}\\
\approx&\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)
\end{aligned}$$

Thus, if we project the mass distribution of the l_p ball of volume 1 onto a single coordinate direction, we get a distribution with density function $f(x)=\exp(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p})=\exp\big(\frac{1}{p}-e(\frac{2}{p}\Gamma(\frac{1}{p}))^{p}x^{p}\big)$.

Thus, for an l_p ball centered at original point O with volume 1 and approximate radius $\frac{d^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}$, then we can use the integral $2\int_0^s\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx$ to estimate the volume between two parallel slices at the same distance s from the center. Then the volume of l_p ball cap corresponding to the slice at distance s from the center can be approximated by $\frac{1}{2}-\int_0^s\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx$. Note that the ratio k of slice's distance d from center to radius r is about $s\Big/\frac{d^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}=\frac{2s(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}{d^{\frac{1}{p}}}$, i.e. $s=\frac{kd^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}$. Thus, the volume of cap can be represented as

$$\frac{1}{2}-\int_0^{\frac{kd^{\frac{1}{p}}}{2(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}}\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx$$

which is only related to the ratio k. Then, we can conclude that for a l_p ball with radius r, when dimension d is large enough and its cap corresponding to the slice at distance h form the center, then the volume ratio of cap to ball is approximately

$$\frac{1}{2}-\int_0^{\frac{sd^{\frac{1}{p}}}{2r(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}}\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx$$

Thus,

$$\frac{\text{Vol}(\Omega_2)}{\text{Vol}(K)}=\frac{1}{2}-\int_0^{\frac{\epsilon d^{\frac{1}{p}}}{4r(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}}\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx$$

and therefore

$$\begin{aligned}
&TV(\mu,\epsilon e_1+\mu)\\
=&1-\frac{\text{Vol}(K\cap(\epsilon e_1+K))}{\text{Vol}(K)}=1-\frac{2\text{Vol}(\Omega_2)}{\text{Vol}(K)}\\
=&2\int_0^{\frac{\epsilon d^{\frac{1}{p}}}{4r(pe)^{\frac{1}{p}}\Gamma(1+\frac{1}{p})}}\exp\Big(\frac{1}{p}-e(2x\Gamma(1+\frac{1}{p}))^{p}\Big)dx
\end{aligned}$$

□

K. Proof of theorem 4.7

Proof. Note that $f(x)$ and $g(x)$ are respectively density functions of reference measure $\rho = x + \mu$ and perturbed measure ν and $q(x)$ is defined as $g(x) - f(x)$. Therefore

$$\begin{aligned}\mathbb{E}_{X\sim\nu}[\phi(X)] &= \int \phi(x)g(x)dx \\ &= \int \phi(x)\big(g(x) - f(x)\big)dx + \int \phi(x)f(x)dx \\ &= \int \phi(x)q(x)dx + \int \phi(x)f(x)dx \\ &= \int \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)]\end{aligned}$$

where the first term contains all the uncertainty in one functional variable $q(x)$ and the second term is a constant when sample point x, smoothing measure μ and specification ϕ are fixed. And when $\nu \in \mathcal{D}_{x,\delta,p}$ or equivalently $W_p(\nu, x+\mu) \le \delta$, applying the dual form of W_p distance given in formula 6 and 7, we have

$$\begin{aligned}W_1(\nu, x+\mu) &= \sup_{\varphi\in\mathcal{F}_1} \int \varphi(x)(f-g)(x)dx \\ &= \sup_{\varphi\in\mathcal{F}_1} \int \varphi(x)(g-f)(x)dx \\ &= \sup_{\varphi\in\mathcal{F}_1} \int \varphi(x)q(x)dx \\ &= \sup_{\|f\|_L\le 1} \int f(x)q(x)dx \le \delta \end{aligned} \tag{51}$$

And when $\nu \in \mathcal{D}_{x,\xi}$ or equivalently $TV(\nu, x+\mu) \le \xi$, applying lemma B.3 for absolutely continuous measure, we have

$$TV(\nu, x+\mu) = \frac{1}{2}\int |f(x) - g(x)|dx = \frac{1}{2}\int |q(x)|dx \le \xi \tag{52}$$

It follows that $OPT(\phi, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$ is equivalent to $\min_{\nu\in\mathcal{D}_{x,\delta,p}\cap\mathcal{D}_{x,\xi}} \mathbb{E}[\phi(X)]$ according to the definition and therefore equivalent to optimization problem 1 which is obviously convex according to 51 and 52. □

L. Proof of theorem 4.8

Recall the following result proved in the section before

$$\mathbb{E}_{X\sim\nu}[\phi(X)] = \int \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)]$$

When $\nu \in \mathcal{D}_{x,\delta,p}$ or equivalently $W_p(\nu, x+\mu) \le \delta$, applying the dual form of W_p distance given in formula 8 and noticing that $\sup_{y\in \mathrm{spt}(\nu)\cup\mathrm{spt}(x+\mu)} \|y\|_2 = \|x\|_2 + R + \max\{\epsilon, \epsilon d^{\frac{1}{2}-\frac{1}{q}}\} := R^*$, we have

$$\begin{aligned}&\left(\sup_{\varphi\in\mathrm{Lip}(p(2R^*)^{p-1})} \int \varphi(y)(g-f)(y)dy - (p-1)(2R^*)^{p-1}\right)^{\frac{1}{p}} \\ &\le W_p(\nu, x+\mu) \le \delta\end{aligned}$$

or equivalently

$$\begin{aligned}&\sup_{\varphi\in\mathrm{Lip}(p(2R^*)^{p-1})} \int \varphi(y)(g-f)(y)dy \\ &= \sup_{\|f\|_L\le p(2R^*)^{p-1}} \int f(x)q(x)dx \le \delta^p + (p-1)(2R^*)^{p-1}\end{aligned} \tag{53}$$

where $\mathrm{Lip}\big(p(2R^*)^{p-1}\big)$ denotes all maps f from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(x) - f(y)| \le p(2R^*)^{p-1}\|x-y\|$ for all $x, y \in K$. Note $OPT(\phi, x+\mu, \mathcal{D}_{x,\delta,p} \cap \mathcal{D}_{x,\xi})$ is equivalent to $\min_{\nu\in\mathcal{D}_{x,\delta,p}\cap\mathcal{D}_{x,\xi}} \mathbb{E}[\phi(X)]$ according to the definition and therefore can be relaxed into optimization problem which is obviously convex according to 53 and 52.

M. Proof of theorem 4.9

Proof. For $0 < p \le 1$, the optimization over q can be solved using Lagrangian duality as follows: we dualize the constraints on q and obtain

$$\begin{aligned}
L(\lambda) =& \inf_{\|q\|_1 \le 2\xi} \left(\int \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \lambda\Big(\sup_{\|f\|_{L,p}\le 1} \int f(x)q(x)dx - \delta\Big)\right) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \inf_{\|q\|_1 \le 2\xi} \sup_{\|f\|_{L,p}\le 1} \left(\int \phi(x)q(x)dx + \lambda\Big(\sup_{\|f\|_{L}\le 1} \int f(x)q(x)dx - \delta\Big)\right) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \inf_{\|q\|_1 \le 2\xi} \sup_{\|f\|_{L,p}\le 1} \int \big(\phi(x) + f(x)\big)q(x)dx - \lambda\delta \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \sup_{\|f\|_{L,p}\le 1} \inf_{\|q\|_1 \le 2\xi} \int \big(\phi(x) + f(x)\big)q(x)dx - \lambda\delta \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \sup_{\|f\|_{L,p}\le 1} \inf_{\|q\|_1 \le 2\xi} - \int \big|\big(\phi(x) + f(x)\big)q(x)\big|dx - \lambda\delta && (54) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - \inf_{\|f\|_{L,p}\le 1} \sup_{\|q\|_1 \le 2\xi} \int \big|\big(\phi(x) + f(x)\big)q(x)\big|dx - \lambda\delta \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\xi \inf_{\|f\|_{L,p}\le 1} \big|\big|\phi(x) + f(x)\big|\big|_\infty - \lambda\delta && (55) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\xi - \lambda\delta && (56)
\end{aligned}$$

where 54 is due to the choice of $q(x)$ such that $\text{sgn}(q(x)) = \text{sgn}(\phi(x) + f(x))$; 55 is due to Holder inequality when $q = 1, p = \infty$; 56 is due to the fact that $\inf_{\|f\|_L \le 1} \big|\big|\phi(x) + f(x)\big|\big|_\infty = 1$ since the range of $\phi(x)$ is $\{\pm 1\}$ in applications and f cannot change suddenly when crossing the decision region boundary of ϕ due to the Lipschitz constant constraint. Similarly, for $p > 1$, we have

$$\begin{aligned}
& L(\lambda) \\
=& \inf_{\|q\|_1 \le 2\xi} \left(\int \phi(x)q(x)dx + \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \lambda\Big(\sup_{\|f\|_L \le p(2R^*)^{p-1}} \int f(x)q(x)dx - \big(\delta^p + (p-1)(2R^*)^{p-1}\big)\Big)\right) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] \\
& + \inf_{\|q\|_1 \le 2\xi} \sup_{\|f\|_L \le p(2R^*)^{p-1}} \left(\int \phi(x)q(x)dx + \lambda\Big(\sup_{\|f\|_L \le p(2R^*)^{p-1}} \int f(x)q(x)dx - \big(\delta^p + (p-1)(2R^*)^{p-1}\big)\Big)\right) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \inf_{\|q\|_1 \le 2\xi} \sup_{\|f\|_L \le p(2R^*)^{p-1}} \int \big(\phi(x) + f(x)\big)q(x)dx - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \sup_{\|f\|_L \le p(2R^*)^{p-1}} \inf_{\|q\|_1 \le 2\xi} \int \big(\phi(x) + f(x)\big)q(x)dx - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] + \sup_{\|f\|_L \le p(2R^*)^{p-1}} \inf_{\|q\|_1 \le 2\xi} - \int \big|\big(\phi(x) + f(x)\big)q(x)\big|dx - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - \inf_{\|f\|_L \le p(2R^*)^{p-1}} \sup_{\|q\|_1 \le 2\xi} \int \big|\big(\phi(x) + f(x)\big)q(x)\big|dx - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\xi \inf_{\|f\|_L \le p(2R^*)^{p-1}} \big|\big|\phi(x) + f(x)\big|\big|_\infty - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big) \\
=& \mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\xi - \lambda\big(\delta^p + (p-1)(2R^*)^{p-1}\big)
\end{aligned}$$

□

N. Proof of theorem 4.10

Recall the certificate by using Hockey-stick divergence provided in table 4 in (Dvijotham et al., 2020a) as below

$$\epsilon_{\text{HS},\beta} \leq \left[\frac{\beta(\theta_a - \theta_b) - |\beta - 1|}{2}\right]_+$$

When $\beta = 1$, it follows that

$$\epsilon_{\text{HS},1} \leq \left[\frac{\theta_a - \theta_b}{2}\right]_+$$

Besides, recall the relaxation radius using Hockey-stick divergence as below

$$\epsilon_{\text{HS},1} = G\Big(\frac{\epsilon}{2\sigma}\Big) - G\Big(-\frac{\epsilon}{2\sigma}\Big) = 2G\Big(\frac{\epsilon}{2\sigma}\Big) - 1$$

And plug it in above inequality, we have

$$\epsilon_{\text{HS},1} = 2G(\frac{\epsilon}{2\sigma}) - 1 \leq \left[\frac{\theta_a - \theta_b}{2}\right]_+$$

And recall the definition of θ_a and θ_b, we have

$$\mathbb{E}_{X\sim x+\mu}[\phi(X)] = \theta_a - \theta_b$$

Thus, our certificate $\mathbb{E}_{X\sim x+\mu}[\phi(X)] - 2\big(2G(\frac{\epsilon}{2\sigma}) - 1\big) \geq 0$ is equivalent to

$$2G(\frac{\epsilon}{2\sigma}) - 1 \leq \frac{\theta_a - \theta_b}{2}$$

Thus, the equivalence relation holds.