Wasserstein Distance Measure Machines

Alain Rakotomamonjy [1] **Abraham Traoré** [1] **Maxime Bérar** [1] **Rémi Flamary** [2] **Nicolas Courty** [3]

Abstract

This paper presents a distance-based discriminative framework for learning with probability distributions. Instead of using kernel mean embeddings or generalized radial basis kernels, we introduce embeddings based on dissimilarity of distributions to some reference distributions denoted as templates. Our framework extends the theory of similarity of Balcan et al. (2008) to the population distribution case and we prove that, for some learning problems, Wasserstein distance achieves low-error linear decision functions with high probability. Our key result is to prove that the theory also holds for empirical distributions. Algorithmically, the proposed approach is very simple as it consists in computing a mapping based on pairwise Wasserstein distances and then learning a linear decision function. Our experimental results show that this Wasserstein distance embedding performs better than kernel mean embeddings and computing Wasserstein distance is far more tractable than estimating pairwise Kullback-Leibler divergence of empirical distributions.

1. Introduction

Most discriminative machine learning algorithms have focused on learning problems where inputs can be represented as feature vectors of fixed dimensions. This is the case of popular algorithms like support vector machines (Schölkopf & Smola, 2002) or random forest (Breiman, 2001). However, there exists several practical situations where it makes more sense to consider input data as set of distributions or empirical distributions instead of a larger collection of single vector. As an example, multiple instance learning (Dietterich et al., 1997) can be seen as learning of a bag of feature vectors and each bag can be interpreted as samples from an underlying unknown distribution. Applications related to political sciences (Flaxman et al., 2015) or astrophysics (Ntampaka et al., 2015) have also considered this learning from distribution point of view for solving some specific machine learning problems. This paper also addresses the problem of learning decision functions that discriminate distributions.

Traditional approaches for learning from distributions is to consider reproducing kernel Hilbert spaces (RKHS) and associated kernels on distributions. In this larger context, several kernels on distributions have been proposed in the literature such as the probability product kernel (Jebara et al., 2004), the Battarachya kernel (Bhattacharyya, 1943) or the Hilbertian kernel on probability measures of Hein & Bousquet (2005). By leveraging on the flurry of distances between distributions (Sriperumbudur et al., 2010), either in a parametric or non parametric way, it is also possible to build definite positive kernel by considering generalized radial basis function kernels of the form

$$K(\mu,\mu') = e^{-\sigma d^2(\mu,\mu')} \tag{1}$$

where μ and μ' are two distributions, $\sigma > 0$ a parameter of the kernel and $d(\cdot,\cdot)$ a distance between two distributions satisfying some appropriate properties so as to make K definite positive (Haasdonk & Bahlmann, 2004).

Another elegant approach for discriminating distributions has been proposed by Muandet et al. (2012). It consists in defining an explicit embedding of a distribution as a mean embedding in a RKHS. Interestingly, if the kernel of the RKHS satisfies some mild conditions then all the information about the distribution is preserved by this mean embedding. Then owing to this RKHS embedding, all the machinery associated to kernel machines can be deployed for learning from these (embedded) distributions.

As we can see, most works in the literature address the question of discriminating distributions by consider either implicit or explicit kernel embeddings. However, is this really necessary? Our observation is that there are many advantages of directly using distances or even dissimilarities between distributions for learning. It would avoid the need for two-stage approaches, computing the distance and then the kernel, as proposed by Póczos et al. (2013) for distribution regression or estimating the distribution and computing the kernel as introduced by Sutherland et al. (2012). Using kernels limits the choice of distribution distances as the

[1] Université Rouen Normandie, LITIS EA4108, [2] Université Cte d'Azur, OCA Lagrange, UMR 7293, CNRS, [3] Université de Bretagne Sud, IRISA, UMR 6074, CNRS. Correspondence to: AR <alain.rakoto@insa-rouen.fr>.

Technical Report, Copyright 2018 by the author(s).

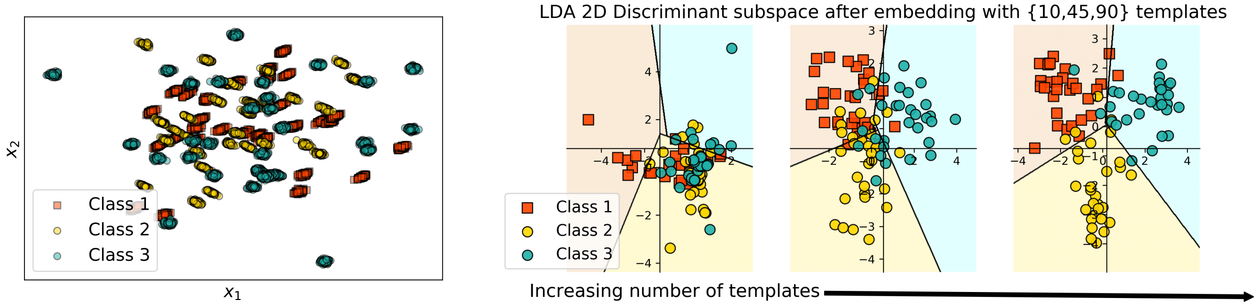

Figure 1. Illustrating the principle of the dissimilarity-based distribution embedding. We want to discriminate empirical normal distributions in $\mathbb{R}^2$; their discriminative feature being the correlation between the two variables. An example of these normal distributions are given in the left panel. The proposed approach consists in computing a embedding based on the dissimilarity of all these empirical distributions (the blobs) to few of, them that serve as templates. Our theoretical results show that if we take enough templates and there is enough samples in each template then with high-probability, we can learn a linear separator that produces few errors. This is illustrated in the 3 other panels in which we represent each of the original distribution as a point after projection in an discriminant 2D space of the embeddings. From left to right, the dissimilarity embedding respectively considers 10, 45 and 90 templates and we can indeed visualize that using more templates improve separability.

resulting kernel has to be definite positive. For instance, Póczos et al. (2012) used Reyni divergences for building generalized RBF kernel that turns out to be non-positive. Finally, computing Gram matrix can be potentially expensive if the number of distribution examples in the training set is large. This work aims at showing that learning from distributions with distances or even dissimilarity is indeed possible. Among all available distances on distributions, we focus our analysis on Wasserstein distances which come with several relevant properties, that we will highlight later, compared to other ones (*e.g* Kullback-Leibler divergence).

Our contributions, depicted graphically in Figure 1, are the following : (I) We show that by following the underlooked works of Balcan et al. (2008), learning to discriminate population distributions with dissimilarity functions comes at no expense. While this might be considered a straightforward extension, we are not aware of any work making this connection. (II) Based on the introduced framework, learnability of a given problem depends on whether the dissimilarity function satisfies some problem-dependent conditions. We show in this paper that, for some problems, the Wasserstein distance satisfies these conditions. (III) Our key theoretical contribution is to show that Balcan's framework also holds for empirical distributions with Wasserstein distance as a dissimilarity function. We provide theoretical guarantees involving a trade-off between the number of distributions and the number of samples in every empirical distribution, in order to get a low-error decision function with high probability. (IV) We illustrate the benefits of using this Wasserstein-based dissimilarity functions compared to kernel approaches in some simulated and real-world vision problems.

2. Framework

In this section, we introduce the global setting and present the theory of learning with dissimilarity functions of Balcan et al. (2008).

2.1. Setting

Define $\mathcal{X}$ as an non-empty subset of $\mathbb{R}^d$ and let $\mathbb{P}$ denotes the set of all probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is σ-algebra of subsets of $\mathcal{X}$. Given a training set $\{\mu_i, y_i\}_{i=1}^n$, where $\mu_i \in \mathbb{P}$ and $y_i \in \{-1, 1\}$, drawn *i.i.d* from a probability distribution P on $\mathbb{P} \times \{-1, 1\}$, our objective is to learn a decision function $h : \mathbb{P} \mapsto \{-1, 1\}$ that predicts the most accurately as possible the label associated to a novel measure μ. In summary, our goal is to learn to classify probability distributions from a supervised setting. While we focus on a binary classification, the framework we consider and analyze can be extended to multi-class classification.

2.2. Dissimilarity function

Most learning algorithms for distributions are based on reproducing kernel Hilbert spaces and leverage kernel value $k(\mu, \mu')$ between two distributions where $k(\cdot, \cdot)$ is the kernel of a given RKHS.

We depart from this approach and instead, we consider learning algorithms that are built from pairwise dissimilarity measures between distribution. The framework we consider is an adaptation of the one proposed by Balcan et al. (2008). Definitions and theorems are reminded and adapted so as to suit our definition of bounded dissimilarity.

Definition 1. a dissimilarity function over $\mathbb{P}$ is any pairwise function $\mathcal{D} : \mathbb{P} \times \mathbb{P} \mapsto [0, M]$.

While this definition emcompasses many functions, given two probability distributions μ and μ', we expect $\mathcal{D}(\mu, \mu')$ to be large when the two distributions are “dissimilar” and to be equal to 0 when they are similar. As such any bounded distance over $\mathbb{P}$ fits into our notion of dissimilarity, eventually after rescaling. Note that unbounded distance which is clipped above M also fits this definition of dissimilarity.

Now, we introduce the definition that characterizes dissimilarity function that allows one to learn a decision function producing low error for a given learning task.

Definition 2. (Balcan et al., 2008) A dissimilarity function $\mathcal{D}$ is a (ϵ, γ)-good dissimilarity function for a learning problem $\mathbf{L}$ if there exists a bounded weighting function w over $\mathbb{P}$, with $w(\mu) \in [0, 1]$ for all $\mu \in \mathbb{P}$, such that a least $1 - \epsilon$ probability mass of distribution examples μ satisfy : $\mathbf{E}_{\mu' \sim P}[w(\mu')\mathcal{D}(\mu, \mu')|\ell(\mu) = \ell(\mu')] + \gamma \leq \mathbf{E}_{\mu' \sim P}\left[w(\mu')\mathcal{D}(\mu, \mu')|\ell(\mu) \neq \ell(\mu')\right]$. The function $\ell(\mu)$ denotes the true labelling function that maps μ to its labels y.

In other words, this definition translates into : a dissimilarity function is “good” if with high-probability, the weighted average of the dissimilarity of one distribution to those of the same label is smaller with a margin γ to the dissimilarity of distributions from the other class.

As stated in a theorem of Balcan et al. (2008), such a good dissimilarity function can be used to define an explicit mapping of a distribution into a space. Interestingly, it can be shown that there exists in that space a linear separator that produces low errors.

Theorem 1. (Balcan et al., 2008) if $\mathcal{D}$ is an (ϵ, γ)-good dissimilarity function, then if one draws a set S from $\mathbb{P}$ containing $n = (\frac{4M}{\gamma})^2 \log(\frac{2}{\delta})$ positive examples $S^+ = \{\nu_1, \cdots, \nu_n\}$ and n negative examples $S^- = \{\zeta_1, \cdots, \zeta_n\}$, then with probability $1 - \delta$, the mapping $\rho_S : \mathbb{P} \mapsto \mathbb{R}^{2n}$ defined as $\rho_S(\mu) = (\mathcal{D}(\mu, \nu_1), \cdots, \mathcal{D}(\mu, \nu_n), \mathcal{D}(\mu, \zeta_1), \cdots, \mathcal{D}(\mu, \zeta_n))$ has the property that the induced distribution $\rho_S(\mathbb{P})$ in $\mathbb{R}^{2n}$ has a separator of error at most $\epsilon + \delta$ at margin at least $\gamma/4$.

The above described framework shows that under some mild conditions on a dissimilarity function and if we consider population distributions, then we can benefit from the mapping ρ_S. We show in the next section that Wasserstein distance respects these conditions for some learning problems. However, in practice, we do have access only to empirical version of these distributions. Our theoretical contribution in Section 4 proves that if the number of distributions n is large enough and enough samples are obtained from each of this distribution, then this framework is applicable with theoretical guarantees to empirical distributions.

3. Wasserstein distance and population learning problem

(ϵ, γ)- goodness of a dissimilarity function is a property that depends on the learning problem. As such, it is difficult to characterize whether a dissimilarity will be good for all problems. In this section, after a brief reminder on the Wasserstein distance, we show that, for some discrimination problems involving normal distributions, this distance satisfies Definition 2 and we exhibit the ϵ and γ associated.

3.1. Brief reminder on Wasserstein distance

Based on the theory of optimal transport, the Wasserstein distance belongs to the class of integral probability metrics that offer means to compare data probability distributions. More formally, we first assume that $\mathcal{X}$ is endowed with a metric $d_{\mathcal{X}}$. Let $p \in (0, \infty)$, and let $\mu \in \mathbb{P}$ and $\nu \in \mathbb{P}$ be two distributions with finite moments of order p (*i.e.* $\int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0)^p d\mu(x) < \infty$ for all x_0 in $\mathcal{X}$). then, the p-Wasserstein distance is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \iint_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}. \tag{2}$$

Here, $\Pi(\mu, \nu)$ is the set of probabilistic couplings π on (μ, ν). As such, for every Borel subsets $A \subseteq \mathcal{X}$, we have that $\mu(A) = \pi(\mathcal{X} \times A)$ and $\nu(A) = \pi(A \times \mathcal{X})$. We refer to (Villani, 2009, Chaper 6) for a complete and mathematically rigorous introduction on the topic. In machine learning, it has recently found numerous applications in important problems such as, for example, multi-label classification (Frogner et al., 2015), domain adaptation (Courty et al., 2017) or generative models (Arjovsky et al., 2017). Its efficiency comes from two major factors: *i)* it handles empirical data distributions without resorting first to parametric representations of the distributions *ii)* the geometry of the underlying space is leveraged through the embedding of the metric $d_{\mathcal{X}}$. Also, there is no restriction on the overlapping of the support of μ and ν, contrary to several alternatives like Kullback-Leibler divergences. In some very specific cases the solution of the infimum problem is analytic. For instance, in the case of two Gaussians $\mu \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\nu \sim \mathcal{N}(\mathbf{m}_2, \Sigma_2)$ the Wasserstein distance with $d_{\mathcal{X}}(x, y) = \|x - y\|_2$ reduces to:

$$W_2^2(\mu, \nu) = ||\mathbf{m}_1 - \mathbf{m}_2||_2^2 + \mathcal{B}(\Sigma_1, \Sigma_2)^2 \tag{3}$$

where $\mathbb{B}(,)$ is the so-called Bures metric (Bures, 1969):

$$\mathcal{B}(\Sigma_1, \Sigma_2)^2 = \text{trace}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}). \tag{4}$$

Yet, in a general machine learning setting, we make no assumption on the form of the distributions, and distributions are observed through samples. In this case, computing the Wasserstein distance boils down to solve a discrete version of Equation 2 which is a linear programming problem. Accelerated solutions of this linear program has been found by adding an entropic regularization term (Cuturi, 2013). In the remainder of the paper we restrict ourselves (without loss of generality or further implications) to the case of the Wasserstein of order 2, and we will note $W_2(\mu,\nu)$ simply as $W(\mu,\nu)$.

3.2. Discriminating normal distributions with the mean

Consider a binary distribution classification problem where samples from both classes are defined by Gaussian distributions in $\mathbb{R}^d$. Means of these Gaussian distribution follow another Gaussian distribution which mean depends on the class while covariance are fixed. Hence, we have $\mu_i \sim \mathcal{N}(\mathbf{m}_i, \Sigma)$ with $\mathbf{m}_i \sim \mathcal{N}(\mathbf{m}^\star_{-1}, \Sigma_0)$ if $y_i = -1$ and $\mathbf{m}_i \sim \mathcal{N}(\mathbf{m}^\star_{+1}, \Sigma_0)$ if $y_i = +1$ where Σ and Σ_0 are some definite-positive covariance matrix. We suppose that both classes have same priors. We also denote $D^\star = \|\mathbf{m}^\star_{-1} - \mathbf{m}^\star_{+1}\|_2^2$ which is a key component in the learnability of the problem. Intuitively, assuming that the volume of each μ_i as defined by the determinant of Σ is smaller than the volume of Σ_0, the larger $D^\star$ is the easier the problem should be. This idea appears formally in what follows.

Based on Wasserstein distance between two normal distributions with same covariance matrix, we have $W(\mu_i,\mu_j)^2 = \|\mathbf{m}_i - \mathbf{m}_j\|_2^2$. In addition, given a μ_i with mean $\mathbf{m}_i$, regardless of its class, we have, with $k \in \{-1,+1\}$:

$$\mathbf{E}_{\mu_j : \mathbf{m}_j \sim \mathcal{N}(\mathbf{m}^\star_k, \Sigma_0)}[\|\mathbf{m}_i - \mathbf{m}_j\|_2^2] = \|\mathbf{m}_i - \mathbf{m}^\star_k\|_2^2 + \mathrm{Tr}(\Sigma_0)$$

Given $\alpha \in]0,1]$, we define the subset of $\mathbb{R}^d$,

$$\mathcal{E}_{-1} = \{\mathbf{m} : (\mathbf{m} - \mathbf{m}_{-1})^\top (\mathbf{m}^\star_{+1} - \mathbf{m}^\star_{-1}) \leq \frac{1-\alpha}{2} D^\star$$

Informally, $\mathcal{E}_{-1}$ is an half-space containing of $\mathbf{m}_{-1}$ for which all points are nearer to $\mathbf{m}_{-1}$ than $\mathbf{m}_{+1}$ with a margin defined by $\frac{1-\alpha}{2} D^\star$. In the same way, we define $\mathcal{E}_{+1}$ as :

$$\mathcal{E}_{+1} = \{\mathbf{m} : (\mathbf{m} - \mathbf{m}^\star_{+1})^\top (\mathbf{m}^\star_{-1} - \mathbf{m}^\star_{+1}) \leq \frac{1-\alpha}{2} D^\star\}$$

Based on these definition, we can state that $W(\cdot,\cdot)$ is a (ϵ,γ) good dissimilarity function with $\gamma = \alpha D^\star$, $\epsilon = \frac{1}{2}\int_{\mathbb{R}^d \setminus \mathcal{E}_{-1}} d\mathcal{N}(\mathbf{m}_{-1}, \Sigma_0) + \frac{1}{2}\int_{\mathbb{R}^d \setminus \mathcal{E}_{+1}} d\mathcal{N}(\mathbf{m}_{+1}, \Sigma_0)$ and $w(\mu) = 1, \forall \mu$. Indeed, it can be shown that for a given μ_i with $y_i = -1$, if $\mathbf{m}_i \in \mathcal{E}_{-1}$ then

$$\|\mathbf{m}_i - \mathbf{m}^\star_{-1}\|_2^2 + \underbrace{\alpha \|\mathbf{m}^\star_{-1} - \mathbf{m}^\star_{+1}\|_2^2}_{\gamma} \leq \|\mathbf{m}_i - \mathbf{m}^\star_{+1}\|_2^2$$

With a similar reasoning, we get an equivalent inequality for μ_i of positive label. Hence, we have all the conditions given in Definition 2 for the Wasserstein distance to be an (ϵ,γ) good dissimilarity function for this problem. Note that the γ and ϵ naturally depend on the distance between expected means. The larger this distance is, the larger the margin and the smaller ϵ are.

Remark 1. While the paper focuses on analyzing Wasserstein distance as a good dissimilarity measure, we can note that for this specific problem of discriminating normal distribution, the Kullback-Leibler divergence defined as:

$$KL(\mu_1,\mu_2) = \int \log\left(\frac{\mu_1}{\mu_2}\right) d\mu_1$$

is also a (ϵ,γ) good dissimilarity function. Indeed, for μ_1 and μ_2 being two normal distribution with same covariance matrix Σ_0, we have $KL(\mu_1,\mu_2) = \|\mathbf{m}_2 - \mathbf{m}_1\|^2_{\Sigma_0^{-1}}$. And following exactly the same steps as above, but replacing inner product $\mathbf{m}^\top \mathbf{m}'$ with $\mathbf{m}^\top \Sigma_0^{-1} \mathbf{m}'$ leads to similar margin $\gamma = \alpha\|\mathbf{m}^\star_{-1} - \mathbf{m}^\star_{+1}\|^2_{\Sigma_0^{-1}}$ and similar definition of ϵ.

3.3. Discriminating normal distributions with the covariance matrix

Consider now a binary distribution classification problem where samples from both classes are defined by Gaussian distributions in $\mathbb{R}^d$ sharing a common mean but with different covariances. We thus can use Equation 3 to compute the squared Wasserstein distance. Let's consider a simple example, where the covariance matrices share a similar structure : constant elements a on the diagonal and a random anti-diagonal element. The distribution of this element is given by $b \sim \mathcal{U}(-a,-a/2)$ if $y_i = -1$ and $b \sim \mathcal{U}(a/2,a)$ if $y_i = +1$. By construction, every matrix shares the same eigenvectors but the association between eigenvalues $\lambda = a \pm b$ and eigenvectors switch between classes. Geometrically classes are distinguished by the orientation of the ellipsis corresponding to the covariances matrices. The greater the quantity $|b/a|$ is (i.e. the flatter the ellipsis are), the easier it is to assign a class.

In such setting, the Wasserstein distance is:

$$W(\mu_i,\mu_j)^2 = \|\mathbf{m}_i - \mathbf{m}_j\|_2^2 + 4a$$
$$- 2\left(\sqrt{(a + y_j b_j)(a + b_i)} + \sqrt{(a - y_j b_j)(a - b_i)}\right).$$

Based on this definition, and following similar steps than in the previous case, we can show (details are in the appendix) that $W(\cdot,\cdot)$ is a (ε,γ)- good dissimilarity function with $\gamma = 2\sqrt{2}\left(\frac{7}{3} - \sqrt{3}\right)\alpha$ and $\epsilon = \frac{1}{2}\int_{[-a,-a/2]\setminus\mathcal{B}_{-1}} \frac{2}{a} db + \frac{1}{2}\int_{[a/2,a]\setminus\mathcal{B}_{+1}} \frac{2}{a} db$ (explicit expression of ϵ can be derived from the equivalent condition).

4. Learning with empirical distributions

In the above sections, we have introduced the notion of (ϵ, γ) good dissimilarity function and we have shown that for some learning problems, involving population normal distributions, the Wasserstein distance satisfies the goodness conditions. Naturally, goodness of Wasserstein distance does not limit to these problems, but proving this property for a larger set of problem is beyond the scope of this paper. Furthermore, in practice, we do not have access to population distributions but to their empirical counterparts. In what follows, we prove that under some conditions on the number of samples in each distribution, it is still possible to learn a separator with low error.

4.1. Theoretical analysis

Suppose that we have at our disposal a dataset composed of $\{\mu_i, y_i = 1\}_{i=1}^n$ where each μ_i is a distribution. However, each μ_i is not observed directly but instead we observe it empirical version $\hat{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\delta_{\mathbf{x}_{i,j}}$ with $\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots \mathbf{x}_{i,N_i} \overset{i.i.d}{\sim} \mu_i$. For a sake of simplicity, we assume in the sequel that the number of samples for all distributions are equal to N.

We now show that, assuming that the Wasserstein distance is a (ϵ, γ) good dissimilarity function on population distributions, for a given learning problem, then building the mapping ρ_S based on empirical distributions still leads to a low-error separator. We formally translate this statement into the following theorem:

Theorem 2. For a given learning problem, if the Wasserstein distance W is an (ϵ, γ)-good dissimilarity function on population distributions, with $w(\mu) = 1,\ \forall \mu$ and K a parameter depending on this dissimilarity then, for a parameter $\lambda \in (0, 1)$, if one draws a set S from $\mathbb{P}$ containing $n = \frac{32M^2}{\gamma^2}\log(\frac{2}{\delta^2(1-\lambda)})$ positive examples $S^+ = \{\nu_1, \cdots, \nu_n\}$ and n negative examples $S^- = \{\zeta_1, \cdots, \zeta_n\}$, and from each distribution ν_i or ζ_i, one draws $N = \frac{256}{K\gamma^2}\log(\frac{1}{\delta^2\lambda})$ samples so as to build empirical distributions $\{\hat{\nu}_i\}$ or $\{\hat{\zeta}_i\}$, then with probability $1 - \delta$, the mapping $\hat{\rho}_S : \mathbb{P} \mapsto \mathbb{R}^{2n}$ defined as

$$\hat{\rho}_S(\hat{\mu}) = \frac{1}{M}(W(\hat{\mu}, \hat{\nu}_1), \cdots, W(\hat{\mu}, \hat{\nu}_n), W(\hat{\mu}, \hat{\zeta}_1), \cdots, W(\hat{\mu}, \hat{\zeta}_n))$$

has the property that the induced distribution $\rho_S(\mathbb{P})$ in $\mathbb{R}^{2n}$ has a separator of error at most $\epsilon + \delta$ and margin at least $\gamma/4$.

Let us point out some relevant insights from this theorem. At first, due to the use of empirical distribution, the sample complexity of the learning problem increases for achieving similar error. Secondly, note that λ has a trade-off role on the number n of samples ν_i and ζ_i and the number of observations per distribution. Hence, for a fixed error $\epsilon+\delta$ at margin $\gamma/4$, having less samples has to be paid by sampling more observations.

The proof of Theorem 3 follows the steps from Balcan et al. but takes advantage of a key technical result on empirical distributions that we present below in Lemma 1. Note that the theorem holds only for the Wasserstein distance because the result in that Lemma applies only to Wasserstein distance. However, we will discuss below under which property it can be extended to a larger class of dissimilarity function.

Lemma 1. Denote μ as a fixed distribution sampled from $\mathbb{P}$ of class y and $\hat{\mu}$ its empirical version composed of N observations. Suppose that we have a set of n distributions $\{\nu_i\}_{i=1}^n$ which have the same label y, then the following concentration inequality holds for any $\epsilon > 0$:

$$\begin{aligned} \Pr \quad & \left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n W(\hat{\mu}, \hat{\nu}_i) - \mathbf{E}_{\nu\sim\mathbb{P}}[W(\mu, \nu)|\ell(\mu) = \ell(\nu)]\right| > \epsilon\right) \\ & \leq e^{-KN\frac{\epsilon^2}{16}} + 2e^{-n\frac{\epsilon^2}{2M^2}} \end{aligned}$$

This lemma tells us that, with high probability, the mean average of the dissimilarity between an empirical distribution and some other empirical distribution *of the same class* does not differ much from the expectation of this dissimilarity measured on population distributions. Interestingly, the bound on the probability is composed of two terms : the first one is related to Wasserstein distance between a distribution and its empirical version while the second one is due to the empirical version of the expectation (resulting thus from Hoeffding inequality). The detailed proof of this result is given in the supplementary material.

Theorem 3 can be extended to any (ϵ, γ) good dissimilarity function under the condition that it satisfies a concentration inequality of the form : Pr $(\mathcal{D}(\mu, \hat{\mu}) > \epsilon) \leq f(N, \epsilon)$ where μ is a population distribution, $\hat{\mu}$ its empirical version with N samples and f is a function of the number of samples in the empirical distribution and of the deviation ϵ, with f going to 0 as N or ϵ goes to ∞.

Hence, from a theoretical point of view, there is only one reason for choosing one (ϵ, γ) good dissimilarity function on population distributions from another. The rationale would be to consider the dissimilarity function with the fastest rate of convergence of the concentration inequality Pr$(\mathcal{D}(\mu, \hat{\mu}) > \epsilon)$, as this rate will impact the upper bound in Theorem 3.

4.2. On the benefit of Wasserstein distance

The approach we advocate for learning with distribution is simple. It consists in computing pairwise Wasserstein distances between all training distributions and some of them denoted as templates, and use this dissimilarity embedding as feature for a linear separator algorithm (such as a SVM). Note that while not theoretically justified by our framework, considering a non-linear classifier such as a Gaussian kernel SVM is also possible and sometimes leads to better results.

Let us now discuss why the Wasserstein distance is an important piece of the proposed approach. From an algorithmic point of view, we need to compute dissimilarities of distributions. There exists several family of distance/divergence between probabilities, most common and popular ones being the ϕ-divergence (Pardo, 2005), for which Kullback-Leibler divergence is a particular case and the integral probability metrics which encompasses the Wasserstein distance. In practice, we need to compute these distances/divergences from samples obtained *i.i.d* from the unknown distribution μ and ν. The problem of estimating in a non-parametric way some ϕ-divergence, especially the Kullback-Leibler divergence have been thoroughly studied by Nguyen et al. (2007; 2010). For KL divergence, these estimations are obtained by solving a quadratic programming problem. In a nutshell, compared to Kullback-Leibler divergence, Wasserstein distance benefits from a linear programming problem compared to a quadratic programming problem, which is far more expensive too compute. In addition, unlike KL-divergence, Wasserstein distance takes into account the properties of $\mathcal{X}$ and as such it does not diverge for distributions that do not share support.

Non-parametric estimation of integral probability metrics has been studied by Sriperumbudur et al. (2010) . In this latter work, they have shown that the Wasserstein distance, with the $d_\mathcal{X}$ being the euclidean norm and $p = 1$ and some other integral probability metrics can be estimated by solving a linear programming problem. For the Wasserstein-1 distance, this was a well-known result based on the Kantorovich-Rubinstein theorem (Villani, 2009). Note that for other generic metrics $d_\mathcal{X}$, the Wassertein distance can be computed, using the discrete version of Equation (2). Interestingly, we shall remark that one integral probability metric, the maximum mean discrepancy (MMD) (Gretton et al., 2007; Sriperumbudur et al., 2010), is a metric between empirical distributions that can be computed in closed-form. MMD is the key component of the Support Measure Machines of Muandet et al. (2012), an extension of Support Vector Machines for discriminating distributions. Compared to another integral probability metrics such as MMD, Wasserstein distance suffers the computational comparison as MMD is just the averaged pairwise kernel evaluation of all samples. Contrarily, Wasserstein distance, by construction, somehow aims at finding the best match between samples and unlike MMD, the resulting distance will depend on fewer pairwise distances between samples of the two distributions. Hence, we believe that the cheapness of computation of MMD comes at the expense of losing ability of finely comparing distributions. But the more interesting argument in favor of Wasserstein is also the ability to compare disributions that do not share support without stagnating as MMD does. This property and the fact that its gradient vanishes except when the distributions are identical have been instrumental to the performances of Wasserstein Generative Adversarial Networks (Arjovsky et al., 2017).

5. Numerical experiments

In this section, we have analyzed and compared the performances of our Wasserstein distances based embedding for learning to classify distributions. Several toy problems, similar to those described in Section 3.2 and 3.3 have been considered as well as a computer-vision real-world problem.

5.1. Competitors

Before describing the experiments we carried out, we first discuss the algorithms we have compared. We have considered two variants of our approach. The first one embeds the distributions based on $\hat{\rho}_S$ by using, unless specified, all distributions available in the training set. Then, we learn either a linear SVM or a Gaussian kernel classifier resulting in two methods dubbed in the sequel as **WDMM + linear SVM** and **WDMM + Gaussian SVM**. In the family of integral probability metrics, we have considered as a competitor the support measure machines of Muandet et al. (2012), denoted as **SMM**. We have considered its non-linear version which used an Gaussian kernel on top of the MMD kernel. In SMM, we have thus two kernel hyperparameters.

Note that in addition to SMM, other kernel on distributions could have been considered by using generalized radial basis function kernel involving squared-distance on distributions as in Equation 1. However, most of these distances can not be computed based on samples and need kernel density estimation (Sutherland et al., 2016). Moreover positive definiteness of resulting kernel may not be guaranteed.

Kullback-Leibler divergence can replace the Wasserstein distance in our framework. For instance, we have highlighted that for the problem in Section 3.2, KL-divergence is an (ϵ, γ) good dissimilarity function. We have thus implemented the non-parametric estimation of the KL-divergence based on quadratic programming (Nguyen et al., 2010; 2007). Pairwise distances have then be used for building an empirical map $\hat{\rho}_S$ followed by a linear or a non-linear classifier, leading to a KL-variant of our approach.

After few experiments on the toy problems, we finally decided to not report performance of the KL-divergence based approach due to its poor computational scalability. For instance, for the first toy problem, KL achieved similar performances than other competitors. However for $n = 200$, it took several days of computations for running 1 trial while other methods had already finished the 20 trials.

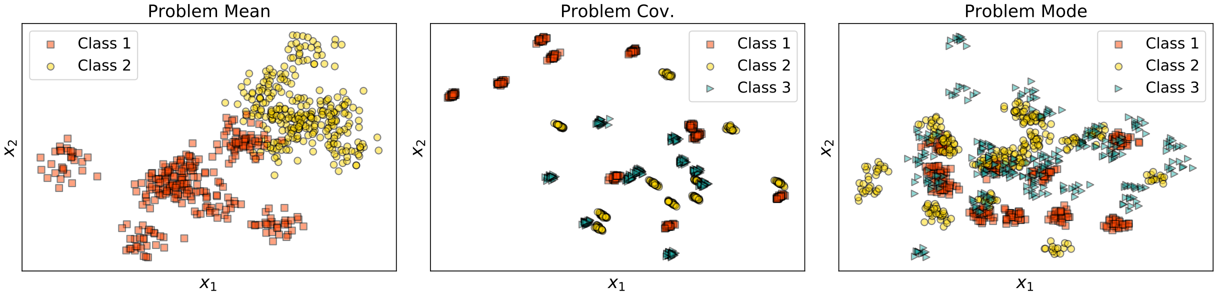

Figure 2. Examples of the 2D toy problems we addressed. For each problem, the number of empirical distributions is 10 per class and we have 30 samples per distribution. (left) binary classification problem as described in Section 3.2. distributions can be discriminated through their means. (middle) 3-class problem in which discriminative features are based on covariance matrix. This is an extension of the problem describe in Section 3.3. (right) 3-class problem based on mixture of Gaussians. Discriminative features are the number of modes in the distributions and their means.

5.2. Simulated problems

These problems aim at studying the performances of our models in controlled setting. The first toy problem is the one describing in Section 3.2 and it also correspond to the toy problem employed by Muandet et al. (2012) In this case, Mean of a given distribution follows a normal distribution which mean is either $\mathbf{m}^\star_{-1} = [1, 1]$ or $\mathbf{m}^\star_{-1} = [-1, -1]$ with an identity covariance matrix. The covariance of the distribution is fixed for the two classes and is $\sigma\mathbf{I}$, with $\sigma = 0.1$. The second toy problem corresponds to the one described in Section 3.2 but with 3 classes. For all classes, mean of a given distribution follows a normal distribution with mean $\mathbf{m}^\star = [1, 1]$ and covariance matrix $\sigma\mathbf{I}$ with $\sigma = 5$. For class i, covariance matrix of a distribution is $\Sigma_i = [a, b_i; b_i, a]$ where $a = 0.005$ with $b_1 \sim U(a/2, a)$, $b_2 \sim U(-a, -a/2)$ and $b_3 = 0$. The last toy aims at discriminating the number of mode in mixture of Gaussians distributions. For all classes, each Gaussian in the mixtures have pre-defined means and covariance matrix $\sigma\mathbf{I}$ with $\sigma = 0.01$. For all classes, sample varies according to a translation defined by a zero-mean unit-variance normal distribution. An example of realization for all toy problems is given in Figure 2.

For these experiments, we have analyzed the effect of the number of training examples n (which is also the number of templates) and the number of samples N in each distribution. We define a trial for a given n, and N as follows. We have sampled the corresponding distributions. We have performed cross-validation on all parameters of all competitors. For WDMM based approaches, this includes the entropic regularization term λ for computing the Wasserstein distance and all classifier parameters. For KL-divergence, we also have a regularization term in the estimator QP problem. For SMM, this involves all kernel and classifier parameters. Approaches are then evaluated on of 2000 distributions. 20 trials have been considered for each n and N.

Figure 3 represents the averaged classification accuracy for $N = 10$ and $N = 30$ samples per classes and with increasing n of the different competitors.

We remark on the left plots of Figure 3 that SMM and WDMMs performs similarly. There is no clear advantage of one method compared to the other regardless of n and N. The problem seems to be easy enough for both approaches. Results for the second toy problem are reported in the middle panels. We note that this problem is more difficult and SMM struggles in achieving good performance. While SMM and WDMM yield similar performances for small n, WDMM benefits from a larger number of training examples. For $n = 1000$, the difference in performance is almost 20% of accuracy. Difference between top and bottom panels also show that both approaches take advantages of the increased number of samples in each distribution. SMM gains about 7% of performances for $n \geq 200$ while WDMM gains approximatively 10%. Interestingly, for this toy problem, using a Gaussian kernel on top of the WD embeddings helps in improving performances. For the third toy problem, we note that WDMM + linear SVM outperforms all competitors. This difference in performance compared to SMM is about 20% for small number of training examples.

5.3. Natural scene categorization

We have compared the performance of SMM and our WDMM approaches on a computer vision problem. For this purpose, we have reproduced the experiments carried out by Muandet et al. (2012). Their idea is to consider an image of a scene as an histogram of codewords, where the codewords have been obtained by k-means clustering of 128-dim SIFT vector and thus to use this histogram as a discrete probability distribution for classifying the images. Details of the feature extraction pipeline can be found in the paper Muandet et al. (2012). The only difference our experimental set-up is that we have used an enriched ver-

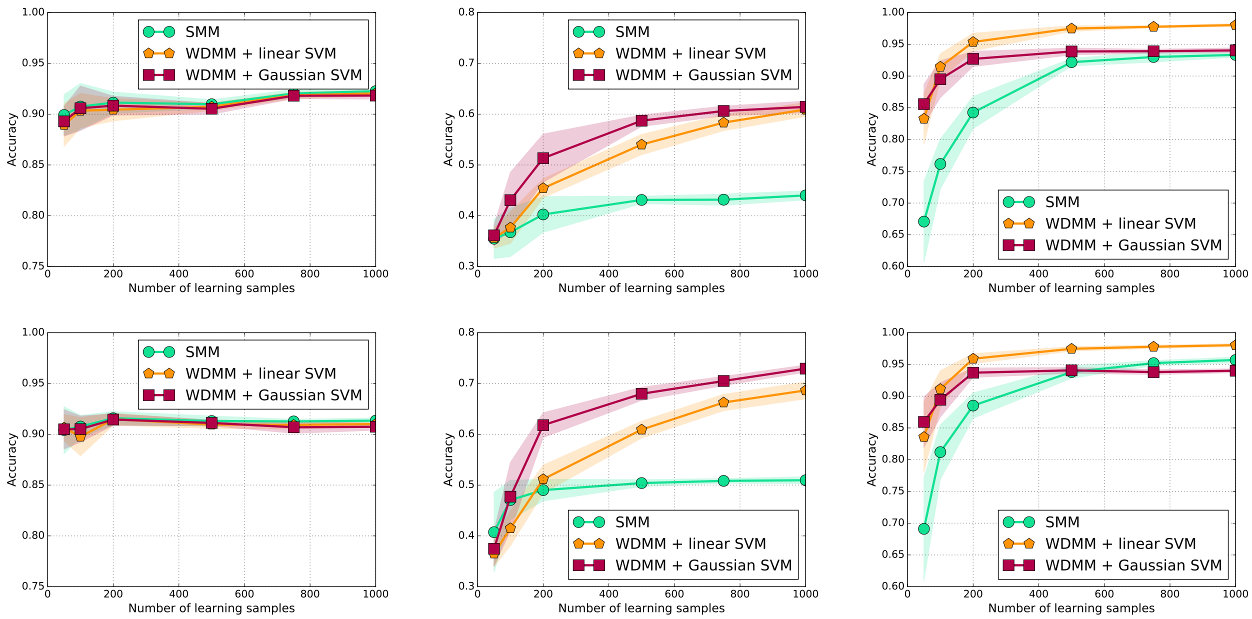

Figure 3. Comparing performances of Support Measure Machines and Wasserstein distance + classifier. From left to right, we have the results on toy problem denoted as **Mean**, **Cov** and **Mod**. The top row represents results when the number of samples per distribution is $N = 10$ for the bottow row $N = 30$. Performances have been averaged over 20 trials.

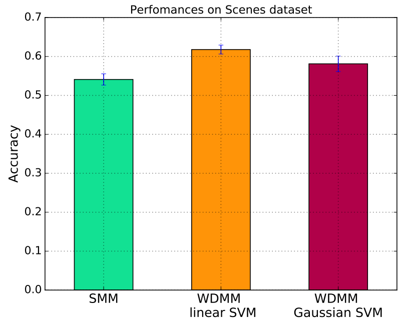

Figure 4. Performances on SMM and WDMM on Scenes dataset. Results are averaged over 10 trials with random drawn of the training set.

sion of the dataset [1] they used. Similarly, we have used 100 images per class for training and the rest for testing. Again, all hyperparameters of all competing methods have been selected by cross-validation.

The averaged results over 10 trials are presented in Figure 4. They illustrate the benefit of WDMM approaches compared to SMM as both linear and non-linear approaches perform better and SMM. We believe that the gain in performance is due to the ability of the Wasserstein distance of matching samples of one distribution to only few samples of the other distribution. By doing so, we believe that it is able to capture in an elegant way complex interaction between samples of distributions.

6. Conclusion

This paper introduces a method for learning to discriminate probability distributions based on dissimilarity functions. The algorithm consists in embedding the distributions into a space of dissimilarity to some template distributions and to learn a linear decision function in that space. From a theoretical point of view, when considering population distributions, our framework is an extension of the one of Balcan et al. (2008). But we provide a theoretical analysis showing that for embeddings based on empirical distributions, given enough samples, we can still learn a linear decision functions with low error with high-probability with empirical Wasserstein distance. The experimental results illustrate the benefits of using empirical dissimilarity on distributions on toy problems and real-world data.

Futur works will be oriented toward analyzing a more general class of regularized optimal transport divergence, such as the Sinkhorn divergence (Genevay et al., 2017) in the context of Wasserstein distance measure machines. Also, we will consider extensions of this framework to regression problems, for which a direct application is not immediate.

[1] The dataset is available at http://www-cvr.ai.uiuc.edu/ponce_grp/data/

References

Arjovsky, Martin, Chintala, Sumit, and Bottou, Léon. Wasserstein generative adversarial networks. In *ICML*, pp. 214–223, 2017.

Balcan, Maria-Florina, Blum, Avrim, and Srebro, Nathan. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

Bhattacharyya, Anil. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

Bolley, François, Guillin, Arnaud, and Villani, Cédric. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

Breiman, Leo. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Bures, Donald. An extension of kakutani's theorem on infinite product measures to the tensor product of semifinite σ-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Courty, Nicolas, Flamary, Rémi, Tuia, Devis, and Rakotomamonjy, Alain. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013.

Dietterich, Thomas G, Lathrop, Richard H, and Lozano-Pérez, Tomás. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89 (1-2):31–71, 1997.

Flaxman, Seth R, Wang, Yu-Xiang, and Smola, Alexander J. Who supported obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 289–298. ACM, 2015.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. Learning with a Wasserstein loss. In *NIPS*. 2015.

Genevay, A., Peyré, G., and Cuturi, M. Learning Generative Models with Sinkhorn Divergences. *ArXiv e-prints*, June 2017.

Gozlan, Nathael. Transport inequalities and concentration of measure. In *ESAIM: Proceedings and Surveys*, pp. 1–23, 2015.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte, Schölkopf, Bernhard, and Smola, Alex J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Haasdonk, Bernard and Bahlmann, Claus. Learning with distance substitution kernels. In *Joint Pattern Recognition Symposium*, pp. 220–227. Springer, 2004.

Hein, Matthias and Bousquet, Olivier. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pp. 136–143, 2005.

Jebara, Tony, Kondor, Risi, and Howard, Andrew. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

Muandet, Krikamol, Fukumizu, Kenji, Dinuzzo, Francesco, and Schölkopf, Bernhard. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pp. 10–18, 2012.

Nguyen, XuanLong, Wainwright, Martin J, and Jordan, Michael I. Nonparametric estimation of the likelihood ratio and divergence functionals. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, pp. 2016–2020. IEEE, 2007.

Nguyen, XuanLong, Wainwright, Martin J, and Jordan, Michael I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Ntampaka, Michelle, Trac, Hy, Sutherland, Dougal J, Battaglia, Nicholas, Póczos, Barnabás, and Schneider, Jeff. A machine learning approach for dynamical mass measurements of galaxy clusters. *The Astrophysical Journal*, 803(2):50, 2015.

Pardo, Leandro. *Statistical inference based on divergence measures*. CRC press, 2005.

Póczos, Barnabás, Xiong, Liang, Sutherland, Dougal J, and Schneider, Jeff. Nonparametric kernel estimators for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2989–2996. IEEE, 2012.

Póczos, Barnabás, Singh, Aarti, Rinaldo, Alessandro, and Wasserman, Larry A. Distribution-free distribution regression. In *AISTATS*, pp. 507–515, 2013.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Sriperumbudur, Bharath K, Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, and Lanckriet, Gert RG.

Non-parametric estimation of integral probability metrics. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pp. 1428–1432. IEEE, 2010.

Sutherland, Dougal J, Xiong, Liang, Póczos, Barnabás, and Schneider, Jeff. Kernels on sample sets via nonparametric divergence estimates. *arXiv preprint arXiv:1202.0302*, 2012.

Sutherland, Dougal J, Oliva, Junier B, Póczos, Barnabás, and Schneider, Jeff G. Linear-time learning on distributions with approximate kernel embeddings. In *AAAI*, pp. 2073–2079, 2016.

Villani, C. *Optimal transport: old and new*. Grund. der mathematischen Wissenschaften. Springer, 2009.

A. All the details of discriminating with the covariance matrix

Consider now a binary distribution classification problem where samples from both classes are defined by Gaussian distributions in $\mathbb{R}^d$ sharing a common mean but with different covariances. Remember that as the covariances differ, the Wasserstein distance between the two normal distributions is now

$$W(\mu_i,\mu_j)^2 = \|\mathbf{m}_i-\mathbf{m}_j\|_2^2+\mathrm{Tr}(\Sigma_i+\Sigma_j-2(\Sigma_i^{1/2}\Sigma_j\Sigma_i^{1/2})^{1/2}).$$

Let's consider a simple example with $p = 2$, where the covariance matrices share a similar structure : constant elements a on the diagonal and a random anti-diagonal element. The distribution of this element is given by $b \sim \mathcal{U}(-a, -a/2)$ if $y_i = -1$ and $b \sim \mathcal{U}(a/2, a)$ if $y_i = +1$. By construction, every matrix shares the same eigenvectors but the association between eigenvalues $\lambda = a \pm b$ and eigenvectors switch between classes. Geometrically classes are distinguished by the orientation of the ellipsis corresponding to the covariances matrices. The greater the quantity $|b/a|$ is (i.e. the flatter the ellipsis are), the easier it is to assign a class.

In such setting, the Wasserstein distance is:

$$\begin{aligned}&W(\mu_i,\mu_j)^2 = \|\mathbf{m}_i-\mathbf{m}_j\|_2^2+4a\\&-2\left(\sqrt{(a+y_jb_j)(a+b_i)}+\sqrt{(a-y_jb_j)(a-b_i)}\right).\end{aligned}$$

The expectation w.r.t. μ_j is then given by

$$\begin{aligned}&\mathbf{E}_{\mu_j}[W(\mu_i,\mu_j)^2] = \|\mathbf{m}_i-\mathbf{m}^\star\|_2^2+Tr(\Sigma_0)+4a\\&-2*2*\frac{2}{3}*\frac{1}{y_j}\sqrt{a}\sqrt{a+b_i}\left((1+y_j)^{3/2}-(1+\frac{y_j}{2})^{3/2}\right)\\&-(-1)\frac{8\sqrt{a}y_j}{3}\sqrt{a-b_i}\left((1-y_j)^{3/2}-(1-\frac{y_j}{2})^{3/2}\right).\end{aligned}$$

ie

$$\begin{aligned}&\mathbf{E}_{\mu_j}[W(\mu_i,\mu_j)^2] = \|\mathbf{m}_i-\mathbf{m}^\star\|_2^2+Tr(\Sigma_0)+4a\\&-\frac{8\sqrt{a}}{3}\sqrt{a+b_i}\left(y_j(1+y_j)^{3/2}-y_j(1+\frac{y_j}{2})^{3/2}\right)\\&-\frac{8\sqrt{a}}{3}\sqrt{a-b_i}\left((-y_j)(1-y_j)^{3/2}-(-y_j)(1-\frac{y_j}{2})^{3/2}\right).\end{aligned}$$

as $y_j(1+y_j) = (1+y_j)$ and $-y_j(1-y_j) = (1-y_j)$

$$\begin{aligned}&\mathbf{E}_{\mu_j}[W(\mu_i,\mu_j)^2] = \|\mathbf{m}_i-\mathbf{m}^\star\|_2^2+Tr(\Sigma_0)+4a\\&-\frac{2\sqrt{2a}}{3}\left(2\sqrt{2}(1+y_j)^{3/2}-y_j(2+y_j)^{3/2}\right)\sqrt{a+b_i}\\&-\frac{2\sqrt{2a}}{3}\left(2\sqrt{2}(1-y_j)^{3/2}-(-y_j)(2-y_j)^{3/2}\right)\sqrt{a-b_i}.\end{aligned}$$

Given $\alpha \in]0,1]$, we define the subset of $[-a, \frac{-a}{2}]$,

$$\mathcal{B}_{-1} = \left\{ b : \sqrt{a-b} \geq \sqrt{a+b} + \sqrt{2a}\alpha \right\}.$$

In geometric term, the selected set consists in the "flatest" ellipsis of the set as the inegality posits some minimal distance between the eigenvalues.

Then, it can be shown that for a given μ_i with $b_i \in \mathcal{B}_{-1}$

$$\mathbf{E}_{\mu_j, y_j=-1}[W(\mu_i,\mu_j)^2] + \gamma \leq \mathbf{E}_{\mu_j, y_j=+1}[W(\mu_i,\mu_j)^2]$$

as shown by the following expression

$$\begin{aligned}
&\mathbf{E}_{\mu_j, y_j=-1}[W(\mu_i,\mu_j)^2] - \mathbf{E}_{\mu_j, y_j=-1}[W(\mu_i,\mu_j)^2] = \\
&- \frac{2\sqrt{2a}}{3}\left(7 - 3\sqrt{3}\right)\left(\sqrt{a-b_i} - \sqrt{a+b_i}\right) \\
&\leq \frac{4}{3}\left(7 - 3\sqrt{3}\right) a\alpha
\end{aligned}$$

In the same way, we define the subset of $[\frac{a}{2}, a]$

$$\mathcal{B}_{+1} = \left\{ b : \sqrt{a+b} \geq \sqrt{a-b} + \alpha\sqrt{2a} \right\}.$$

and with a similar reasoning, we get similar inequality for μ_i of positive label. Based on these definition, we can state that $\mathbf{W}(\dot{;})$ is a (ε, γ)- good dissimilarity function with $\gamma = 2\sqrt{2}\left(\frac{7}{3} - \sqrt{3}\right)\alpha$ and $\epsilon = \int_{[-a,-a/2]\setminus\mathcal{B}_{-1}} 1/adb + \int_{[a/2,a]\setminus\mathcal{B}_{+1}} 1/adb$ (explicit expression of ϵ can be derived from the equivalent condition $|b| \geq \alpha\sqrt{2-\alpha^2 a}$).

B. Preliminary results

B.1. Property I

Let W be the Wasserstein distance on $P \times P$ Let μ_i, μ_j be two probability distribution and $\hat{\mu}_i, \hat{\mu}_j$ their empirical version. We have:

$$|W(\mu_i,\mu_j) - W(\hat{\mu}_i,\hat{\mu}_j)| \leq W(\mu_i,\hat{\mu}_i) + W(\mu_j,\hat{\mu}_j) \quad (5)$$

Proof. Owing to triangular inequality, we have

$$\begin{aligned}
W(\mu_i,\mu_j) &\leq W(\mu_i,\hat{\mu}_i) + W(\hat{\mu}_i,\mu_j) & (6)\\
&\leq W(\mu_i,\hat{\mu}_i) + W(\hat{\mu}_i,\hat{\mu}_j) + W(\hat{\mu}_j,\mu_j) & (7)
\end{aligned}$$

We thus have

$$W(\mu_i,\mu_j) - W(\hat{\mu}_i,\hat{\mu}_j) \leq W(\mu_i,\hat{\mu}_i) + W(\hat{\mu}_j,\mu_j)$$

In addition, we also have:

$$\begin{aligned}
W(\hat{\mu}_i,\hat{\mu}_j) &\leq W(\hat{\mu}_i,\mu_i) + W(\mu_i,\hat{\mu}_j) & (8)\\
&\leq W(\hat{\mu}_i,\mu_i) + W(\mu_i,\mu_j) + W(\mu_j,\hat{\mu}_j) & (9)
\end{aligned}$$

Hence :

$$-(W(\mu_i,\mu_j) - W(\hat{\mu}_i,\hat{\mu}_j)) \leq W(\hat{\mu}_i,\mu_i) + W(\mu_j,\hat{\mu}_j) \quad (10)$$

leading to

$$|W(\mu_i,\mu_j) - W(\hat{\mu}_i,\hat{\mu}_j)| \leq W(\hat{\mu}_i,\mu_i) + W(\mu_j,\hat{\mu}_j)$$

□

B.2. Property II

This result is a direct application of concentration inequality of Wasserstein distance (Bolley et al., 2007) applied to Gaussian distribution and to distribution defined on compact space.

Lemma 2. Let μ_i be a probability distribution on a metric space $(\mathbb{R}^n, d)$, $\hat{\mu}_i$ the associated empirical distribution and N the number of samples. If μ_i is a Gaussian distribution or μ_i has a compact support, $\exists K > 0, \mathbb{P}(W(\mu_i,\hat{\mu}_i) > \epsilon) \leq e^{-KN\epsilon^2}$

Proof. **First case**: if μ_i is a Gaussian distribution on $\mathbb{R}^n$, it is well known that the Talagrand inequality $\mathbf{T}_2$ is verified (see (Gozlan, 2015): *Theorem 2.3*). Hence, the property is a direct application of the theorem 1.1 in (Bolley et al., 2007). Remark: roughly speaking, the Talagrand inequality is defined by $W(\mu,\nu) \leq \sqrt{\frac{2}{\lambda}\mathbf{H}(\nu|\mu)}, \mathbf{H}$ being the relative entropy of ν with respect to μ(for more information on Talagrand inequalities, see (Bolley et al., 2007): section 1.1)

Second case: if μ_i has a compact support $\mathcal{K} \subset \mathbb{R}^n$, we have $\forall y \in \mathbb{R}^n$ and $\forall \alpha > 0$:

$$\begin{aligned}
\int_{\mathbb{R}^n} e^{\alpha d^2(x,y)} d\mu(x) &= \int_{\mathbb{R}^n} e^{\alpha d^2(x,y)} d\mu(x)\\
&\leq \max_{x\in\mathcal{K}}(d^2(x,y))\mu(\mathcal{K})\\
&\leq \max_{x\in\mathcal{K}}(d^2(x,y))
\end{aligned}$$

The last inequality is due to the fact that μ is a probability distribution. Since a distance is a continuous function and a continuous function attains its maximum on a compact set, we have:
$\int_{\mathbb{R}^n} e^{\alpha d^2(x,y)} d\mu(x) < \infty$: this corresponds to the existence of a square-exponential moment for the probability distribution μ, which implies the Talagrand inequality $\mathbf{T}_1$ (see (Bolley et al., 2007)). With this condition, the theorem 1.1 in (Bolley et al., 2007) gives our property. □

B.3. Property III

Given a distribution μ_i, if the distributions $\{\mu_j\}_{1\leq j\leq n}$ are independent, the following inequality holds:

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j) - \mathbb{E}_{\mu\sim\mathbb{P}}(W(\mu_i, \mu))| > \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{M^2}}$$

Proof. Since we consider a bounded Wasserstein distance W by a positive constant M, the random variables$\{W(\mu_i, \mu_j)\}_{1\leq j\leq n}$ are bounded by M with probability 1. Given that the variables $\{W(\mu_i, \mu_j)\}_{1\leq j\leq n}$ are bounded (boundedness of W) and independent (due the independence of the $\{\mu_j\}_{1\leq j\leq n}$, the continuity of W and its boundedness) , the *Hoeffding's inequality* yields:

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j) - \mathbb{E}_{\mu\sim\mathbb{P}}(W(\mu_i, \mu))| > \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{M^2}}$$

□

B.4. Property IV

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two subsets of a set Ω.
$\mathbf{S}_1 \cap \mathbf{S}_2 = \{x \in \Omega | x \in \mathbf{S}_1, x \in \mathbf{S}_2\} = \emptyset \Rightarrow \mathbf{S}_1 \subset \Omega \backslash \mathbf{S}_2 = \{x \in \Omega | x \notin \mathbf{S}_2\}$
Proof:
Let $x \in \mathbf{S}_1$. There are two possibilities: $\mathbf{x} \in \mathbf{S}_1 \setminus \mathbf{S}_2$ or $\mathbf{S}_1 \cap \mathbf{S}_2$. Since $\mathbf{S}_1 \cap \mathbf{S}_2 = \emptyset$ by hypothesis, we have:
$x \in \mathbf{S}_1 \setminus \mathbf{S}_2$
$\Rightarrow x \in \mathbf{S}_1$ and $x \notin \mathbf{S}_2$
$\Rightarrow x \in \Omega$ and $x \notin \mathbf{S}_2$ (since $\mathbf{S}_1 \subset \Omega$)
$\Rightarrow x \in \Omega \setminus \mathbf{S}_2$

C. Proof of Lemma 1 in the paper

Our main result is the inequality:

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} W(\hat{\mu}_i, \hat{\mu}_j) - \mathbb{E}_{\mu_j\sim\mathbb{P}}(W(\mu_i, \mu_j))| > \epsilon) \leq g(\epsilon)$$

with $g(\epsilon) = e^{-\frac{KN}{16}\epsilon^2} + 2e^{-\frac{n\epsilon^2}{2M^2}}, \forall\epsilon > 0$, $\hat{\mu}_j$ referring to the empirical distribution of μ_j.
One can easily notice that $g(\epsilon) \to 0$ when $\epsilon \to \infty$.

Proof. Let's denote

$$\Gamma = |\frac{1}{n}\sum_{j=1}^{n} W(\hat{\mu}_i, \hat{\mu}_j) - \mathbb{E}_{\mu_j\sim\mathbb{P}}(W(\mu_i, \mu_j))|$$

$$\Gamma = |\frac{1}{n}\sum_{j=1}^{n} W(\hat{\mu}_i, \hat{\mu}_j) - \frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j) + \frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j) \\ - \mathbb{E}_{\mu_j\sim\mathbb{P}}(W(\mu_i, \mu_j))|$$

By triangular inequality for the absolute value, we have:

$$\Gamma \leq \Gamma_2 + \Gamma_1 \tag{11}$$

with:

$$\Gamma_2 = |\frac{1}{n}\sum_{j=1}^{n} W(\hat{\mu}_i, \hat{\mu}_j) - \frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j)|$$

$$\Gamma_1 = |\frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \mu_j) - \mathbb{E}_{\mu_j\sim\mathbb{P}}(W(\mu_i, \mu_j))|$$

We now have $\Gamma_2 \leq \frac{1}{n}\sum_{j=1}^{n} |W(\hat{\mu}_i, \hat{\mu}_j) - W(\mu_i, \mu_j))|$ (triangular inequality for the absolute value) and

$$\Gamma_2 \leq \frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \hat{\mu}_i) + W(\mu_j, \hat{\mu}_j)$$

owing to the following preliminary results section: **Property I**)
The set $\{W(\mu_j, \hat{\mu}_j)\}_{1\leq j\leq n}$ is a finite set of real numbers. Thus, it admits at least one maximum. Let's denote n_0 the index of one maximum, i.e.:
$n_0 = argmax_{1\leq j\leq n} W(\mu_j, \hat{\mu}_j)$
The last implication (obtained by the property I in the preliminary results section) yields:
$\Gamma_2 \leq \frac{1}{n}\sum_{j=1}^{n} W(\mu_i, \hat{\mu}_i) + W(\mu_j, \hat{\mu}_j)$
$\Rightarrow \Gamma_2 \leq \frac{1}{n}\sum_{i=1}^{n} 2W(\mu_{n_0}, \hat{\mu}_{n_0})$ (by definition of n_0)
$\Rightarrow \Gamma_2 \leq 2W(\mu_{n_0}, \hat{\mu}_{n_0})$
$\Rightarrow \mathbb{P}(W(\mu_{n_0}, \hat{\mu}_{n_0}) \leq \epsilon_1) \leq \mathbb{P}(\Gamma_2 \leq 2\epsilon_1)$ (because $W(\mu_{n_0}, \hat{\mu_{n_0}}) \leq \epsilon_1 \Rightarrow \Gamma_2 \leq 2\epsilon_2$)
$\Rightarrow 1 - \mathbb{P}(W(\mu_{n_0}, \hat{\mu}_{n_0}) \leq \epsilon_1) \geq 1 - \mathbb{P}(\Gamma_2 \leq 2\epsilon_1)$
$\Rightarrow \mathbb{P}(W(\mu_{n_0}, \hat{\mu}_{n_0}) > \epsilon_1) \geq \mathbb{P}(\Gamma_2 > 2\epsilon_1)$.
Since $\mathbb{P}(W(\mu_{n_0}, \hat{\mu}_{n_0}) > \epsilon_1) \leq e^{-KN\epsilon_1^2}$ (property II of the preliminary results section), we have:
$\mathbb{P}(\Gamma_2 > 2\epsilon_1) \leq \mathbb{P}(W(\mu_{n_0}, \hat{\mu}_{n_0}) > \epsilon_1) \leq e^{-KN\epsilon_1^2}$
$\Rightarrow 1 - \mathbb{P}(\Gamma_2 \leq 2\epsilon_1) \leq e^{-KN\epsilon_1^2}$
$\Rightarrow 1 - e^{-KN\epsilon_1^2} \leq \mathbb{P}(\Gamma_2 \leq 2\epsilon_1)$
$\Rightarrow 1 - e^{-KN\epsilon_1^2} \leq \mathbb{P}(-\Gamma_2 \geq -2\epsilon_1)$
Thus:

$$1 - e^{-KN\epsilon_1^2} \leq \mathbb{P}(\epsilon - \Gamma_2 \geq \epsilon - 2\epsilon_1), \forall\epsilon_1 > 0, \epsilon > 0 \tag{12}$$

By definition, we have:
$\Gamma \leq \Gamma_2 + \Gamma_1$(see the equation (11))
$\mathbb{P}(\epsilon \leq \Gamma) \leq \mathbb{P}(\epsilon \leq \Gamma_1 + \Gamma_2)$ (because $\epsilon \leq \Gamma \Rightarrow \epsilon \leq \Gamma_1 + \Gamma_2$)
Hence, we have:

$$\mathbb{P}(\epsilon \leq \Gamma) \leq \mathbb{P}(\epsilon - \Gamma_2 \leq \Gamma_1) \tag{13}$$

At this point, we have the two equations given by (12) and (13):

$\mathbb{P}(\epsilon \leq \Gamma) \leq \mathbb{P}(\epsilon - \Gamma_2 \leq \Gamma_1)$ and $1 - e^{-KN\epsilon_1^2} \leq \mathbb{P}(\epsilon - \Gamma_2 \geq \epsilon - 2\epsilon_1), \forall \epsilon_1 > 0, \epsilon > 0$.

By choosing $\epsilon_1 = \frac{\epsilon}{4}$, we have:

$$\mathbb{P}(\epsilon \leq \Gamma) \leq \mathbb{P}(\underbrace{\epsilon - \Gamma_2 \leq \Gamma_1}_{\mathbf{S}_1}) \tag{14}$$

and

$$1 - e^{-\frac{KN\epsilon^2}{16}} \leq \mathbb{P}(\underbrace{\epsilon - \Gamma_2 \geq \frac{\epsilon}{2}}_{\mathbf{S}_2}), \forall \epsilon > 0 \tag{15}$$

- First Case: $\mathbf{S}_1 \cap \mathbf{S}_2 \neq \emptyset$

$\mathbf{S}_1 = (\mathbf{S}_1 \setminus \mathbf{S}_2) \cup (\mathbf{S}_1 \cap \mathbf{S}_2)$

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq \mathbb{P}(\mathbf{S}_1 \setminus \mathbf{S}_2) + \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2)$

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq \mathbb{P}(\Omega \setminus \mathbf{S}_2) + \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2)$ $(\mathbf{S}_1 \setminus \mathbf{S}_2 \subset \Omega \setminus \mathbf{S}_2, \Omega$ being the universe)

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq 1 - \mathbb{P}(\mathbf{S}_2) + \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2)$(**)

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq e^{-\frac{KN\epsilon^2}{16}} + \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2)$ (by equation (12))

$\mathbf{S}_1 \cap \mathbf{S}_2 \subset \left\{\frac{\epsilon}{2} \leq \Gamma_1\right\}$

$\Rightarrow \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2) \leq \mathbb{P}(\Gamma_1 \geq \frac{\epsilon}{2})$

Given that $\mathbb{P}(\mathbf{S}_1) \leq 1 - \mathbb{P}(\mathbf{S}_2) + \mathbb{P}(\mathbf{S}_1 \cap \mathbf{S}_2)$ by (**), we have by the equation (15):

$\mathbb{P}(\mathbf{S}_1) \leq e^{-\frac{KN\epsilon^2}{16}} + \mathbb{P}(\Gamma_1 \geq \frac{\epsilon}{2})$

Since $\mathbb{P}(\epsilon \leq \Gamma) \leq \mathbb{P}(\mathbf{S}_1)$ by the equation (14), we have:

$\Rightarrow \mathbb{P}(\epsilon \leq \Gamma) \leq e^{-\frac{KN\epsilon^2}{16}} + \mathbb{P}(\Gamma_1 \geq \frac{\epsilon}{2}), \forall \epsilon > 0$

$\Rightarrow \mathbb{P}(\epsilon \leq \Gamma) \leq e^{-\frac{KN}{16}\epsilon^2} + \mathbb{P}(\Gamma_1 \geq \frac{\epsilon}{2}) \leq e^{-\frac{KN}{16}\epsilon^2} + 2e^{-\frac{n\epsilon^2}{2M^2}}$ (property III of the preliminary results section)

- Second case: $\mathbf{S}_1 \cap \mathbf{S}_2 = \emptyset$

By the property IV of the preliminary results section, we have:

$\mathbf{S}_1 \subset \Omega \setminus \mathbf{S}_2$

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq \mathbb{P}(\Omega \setminus \mathbf{S}_2) = 1 - \mathbb{P}(\mathbf{S}_2)$

$\Rightarrow \mathbb{P}(\mathbf{S}_1) \leq e^{-\frac{KN\epsilon^2}{16}}, \forall \epsilon > 0$ (by the equation (15))

$\Rightarrow \mathbb{P}(\Gamma \geq \epsilon) \leq e^{-\frac{KN\epsilon^2}{16}}, \forall \epsilon_1, \epsilon > 0$

$\Rightarrow \mathbb{P}(\Gamma \geq \epsilon) \leq e^{-\frac{KN}{16}\epsilon^2} + \mathbb{P}(\Gamma_1 \geq \frac{\epsilon}{2})$

$\Rightarrow \mathbb{P}(\Gamma \geq \epsilon) \leq e^{-\frac{KN}{16}\epsilon^2} + 2e^{-\frac{n\epsilon^2}{2M^2}}, \forall \epsilon > 0$ (property III of the preliminary results section)

□

D. Proof of Theorem 2

Theorem 3. For a given learning problem, if the Wasserstein distance W is an (ϵ, γ)-good dissimilarity function on population distributions, with $w(\mu) = 1, \ \forall \mu$, and K a parameter depending on this dissimilarity then, for a parameter $\lambda \in (0, 1)$, if one draws a set S from $\mathbb{P}$ containing

$$n = \frac{32M^2}{\gamma^2} \log(\frac{2}{\delta^2(1-\lambda)})$$

positive examples $S^+ = \{\nu_1, \cdots, \nu_n\}$ and n negative examples $S^- = \{\zeta_1, \cdots, \zeta_n\}$, and from each distribution ν_i or ζ_i, one draws

$$N = \frac{256}{K\gamma^2} \log(\frac{1}{\delta^2 \lambda})$$

samples so as to build empirical distributions $\{\hat{\nu}_i\}$ or $\{\hat{\zeta}_i\}$, then with probability $1 - \delta$, the mapping $\hat{\rho}_S : \mathbb{P} \mapsto \mathbb{R}^{2n}$ defined as

$$\hat{\rho}_S(\hat{\mu}) = \frac{1}{M}(W(\hat{\mu}, \hat{\nu}_1), \cdots, W(\hat{\mu}, \hat{\nu}_n), W(\hat{\mu}, \hat{\zeta}_1), \cdots, W(\hat{\mu}, \hat{\zeta}_n))$$

has the property that the induced distribution $\rho_S(\mathbb{P})$ in $\mathbb{R}^{2n}$ has a separator of error at most $\epsilon + \delta$ at margin at least $\gamma/4$. □

Proof. Note that we reproduce here the proof of Balcan et al but we invoke Lemma 1 of the paper at some points.

Consider the linear separator $\tilde{w}$ in the $\hat{\rho}_S$ space defined as $\tilde{w}_i = 1$, for $i \in \{1, \cdots, n\}$ and $\tilde{w}_i = -1$, for $i \in \{n+1, \cdots, 2n\}$. We will show that, with probability at least $(1 - \delta)$, $\tilde{w}$ has error at most $\epsilon + \delta$ at margin $\gamma/4$. Let **Good** be the set of μ satisfying the inequality $\mathbf{E}_{\mu' \sim P}[w(\mu')W(\mu, \mu')|\ell(\mu) = \ell(\mu')] + \gamma \leq \mathbf{E}_{\mu' \sim P}[w(\mu')W(\mu, \mu')|\ell(\mu) \neq \ell(\mu')]$. By assumption, we have $\Pr_{\mu \in P}[\mu \in \text{Good}] \geq 1 - \epsilon$.

Consider some fixed point $\mu \in$ Good. We show that for any such μ,

$$\Pr_{S^+, S^-}\left(\ell(\mu)\frac{\tilde{w}^\top \hat{\rho}_S(\hat{\mu})}{\|\tilde{w}^\top\| \|\hat{\rho}_S(\hat{\mu})\|} \geq \frac{\gamma}{4}\right) \geq 1 - \delta^2.$$

To do so, based on Lemma 1, we notice that n and N are large enough so that with high probability, at least $1 - \delta^2$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} W(\hat{\mu}, \hat{\nu}_i) - \mathbf{E}_{\nu \sim \mathbb{P}}[W(\mu, \nu)|\ell(\mu) = \ell(\nu)]\right| \leq \frac{\gamma}{4}$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n} W(\hat{\mu}, \hat{\zeta}_i) - \mathbf{E}_{\zeta \sim \mathbb{P}}[W(\mu, \zeta)|\ell(\mu) = \ell(\zeta_i)]\right| \leq \frac{\gamma}{4}$$

Let's consider now the case when $\ell(\mu) = 1$. In this case, we have $\ell(\mu)\tilde{w}^\top \hat{\rho}_S(\hat{\mu}) = \frac{n}{M}\left(\frac{1}{n}\sum_{i=1}^n w(\nu_i)W(\hat{\mu},\hat{\nu}_i) - \frac{1}{n}\sum_{i=1}^n w(\zeta_i)W(\hat{\mu},\hat{\zeta}_i)\right)$, and so combining these facts, we have that with probability at least $(1-\delta^2)$ the following holds : $\ell(\mu)\tilde{w}^\top \hat{\rho}_S(\hat{\mu}) \geq \frac{n}{M}(\mathbf{E}_{\nu\sim\mathbb{P}}[w(\nu)W(\mu,\nu)|\ell(\nu) = 1] - \gamma/4 - \mathbf{E}_{\nu\sim\mathbb{P}}[w(\nu)W(\mu,\nu)|\ell(\nu) = -1] - \gamma/4)$. Since $\mu \in$ Good, this then implies that $\ell(\mu)\tilde{w}^\top \hat{\rho}_S(\hat{\mu}) \geq \frac{n}{M}\gamma/2$. Finally, since $\tilde{w}(\nu) \in [-1,1]$ for all ν and since $W(\mu,\nu) \in [0,M]$ for all pairs ν,μ, we have $\|\tilde{w}\| \leq \sqrt{2n}$ and $\|\rho_S\| \leq \frac{1}{M}\sqrt{2n}$ which implies that

$$\mathrm{Pr}_{S^+,S^-}\left(\ell(\mu)\frac{\tilde{w}^\top \hat{\rho}_S(\mu)}{\|\tilde{w}^\top\|\|\hat{\rho}_S(\mu)\|} \geq \frac{\gamma}{4}\right) \geq 1-\delta^2.$$

Since the above holds for any $\mu \in$ Good, it is also true for random $\mu \in$ Good, which implies by Markov's inequality that with probability $1-\delta$, the vector $\tilde{w}$ has error at most δ at margin $\gamma/4$ over P restricted to distributions $\mu \in$ Good. Adding back the ϵ probability mass of points μ not satisfying $\mathbf{E}_{\mu'\sim P}[w(\mu')\mathcal{D}(\mu,\mu')|\ell(\mu) = \ell(\mu')] + \gamma \leq \mathbf{E}_{\mu'\sim P}\left[w(\mu')\mathcal{D}(\mu,\mu')|\ell(\mu) \neq \ell(\mu')\right]$ yields the theorem. □