A survey for the two sample problem

Peyman Morteza
peyman@cs.wisc.edu

Abstract

In this note, we give a survey of results for the two sample problem with focus on the approach that uses kernel theory methods. We focus on the results with provable guarantees and discuss recent developments.

1 Roadmap

The organization of this note would be as follows. In Section 2, we set the notation that we will be using throughout the note and formulate the main problem of study. We also recall facts and notions that we will need to discuss the results in the later sections. In Section 3, we discuss designing hypothesis tests motivated by maximum mean discrepancy and also discuss some recent developments that leads to obtaining efficient tests with improved power. In Section 4, we discuss different proposed techniques for selecting a kernel for the statistical testing procedure. Finally, in Section 5, we discuss applications of the two sample problem and the methods explained in this note in various other domains.

2 Notation, background and problem formulation

2.1 Notation

Throughout this note $\mathcal{X}$ denotes a Polish metric space. Unless otherwise is noted, we assume that $\mathcal{X}$ is compact. Let $d_{\mathcal{X}}$ denotes the metric of $\mathcal{X}$ and Σ denotes the σ-algebra generated by Borel sets of $\mathcal{X}$ and $P(\mathcal{X}, \Sigma)$ denotes the set of all probability measures on $(\mathcal{X}, \Sigma)$. For $p \in P(\mathcal{X}, \Sigma)$, $X_m^p := \{x_1^p, ..., x_m^p\}$ denote an i.i.d sample from p of size m. Also, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denotes a kernel on $\mathcal{X}$. Unless otherwise noted, we assume that kernels are at least continuous, bounded and positive definite.

2.2 The two sample problem

Let $p, q \in P(\mathcal{X}, \Sigma)$ and $X_m^p = \{x_1^p, ..., x_m^p\}$ and $X_n^q = \{x_1^q, ..., x_n^q\}$ be two i.i.d samples from p and q of size m and n respectively. The main question we study in this note is the following:

Given X_m^p and X_n^q, how can one decide whether $p = q$ or not?

We next recall some notions from statistical hypothesis testing framework that we will need to analyze the results related to the two sample problem.

2.3 Statistical hypothesis testing

Here we briefly recall some basic notions from statistical hypothesis testing framework and we refer to [5] for a detailed discussion. For $p, q \in P(\mathcal{X}, \Sigma)$, we are interested in testing the null hypothesis of $p = q$ against the alternative hypothesis $p \neq q$. Formally we write,

$$H_0 : p = q,$$
$$H_1 : p \neq q.$$

Preprint. Under review.

To test above hypothesis, one need a *test statistic* that is computed using samples X_m^p and X_n^q. More formally, a test statistics can described by a mapping,

$$T : \mathcal{X}^m \times \mathcal{X}^n \to \mathbb{R}.$$

A *statistical hypothesis testing procedure*, τ, can be thought of as a mapping,

$$\tau : \mathcal{X}^m \times \mathcal{X}^n \to \{\text{reject}, \text{accept}\},$$

where,

$$\tau(X_m^p, X_n^q) = \begin{cases} \text{reject} & T(X_m^p, X_n^q) \geq \gamma \\ \text{accept} & T(X_m^p, X_n^q) < \gamma \end{cases}.$$

We refer to γ in the above as the *design parameter* of the test. Next, there are two ways the testing procedure can make a mistake,

$$\text{I: } \tau \text{ rejects } H_0 \text{ but } H_0 \text{ is true,}$$
$$\text{II: } \tau \text{ accepts } H_0 \text{ but } H_0 \text{ is false.}$$

We refer to the above as *Type I error* and *Type II error* respectively. The *power of* τ is defined to be the probability of avoiding the Type II error. More specifically,

$$\text{power of } \tau := \mathbb{P}(\tau \text{ rejects } H_0 | H_0 \text{ is false}).$$

Next, we say a statistical test τ is *consistent* if its power approches to 1 as the total number of samples approaches to infinity. Finally, the *level of* τ is defined to be a number $0 \leq \alpha \leq 1$ such that,

$$\mathbb{P}(\tau \text{ rejects } H_0 | H_0 \text{ is true}) \leq \alpha,$$

In other words, the level of the test bounds the probability of Type I error.

2.4 Kernel theory

Here we recall some notions from kernel theory that we will need. We refer to [26] and [24] for detailed discussion about kernel theory. Let $\mathcal{X}$ denotes the background metric space which we will think of it as feature space. First, consider a *feature mapping* ϕ,

$$\phi : \mathcal{X} \to \mathcal{H},$$

where $\mathcal{H}$ is some Hilbert space. It follows from kernel theory that all "relevant information" that is needed from the feature map ϕ is stored the kernel associated to it that is defined by,

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$k(x, y) = \langle \phi(x), \phi(y) \rangle.$$

Conversely, one can start from a kernel function k and reverse the process to construct a Hilbert space $\mathcal{H}_k$ associated to it. First, consider the following vector space,

$$\mathcal{H} := \{f | f : \mathcal{X} \to \mathbb{R}\},$$

and the feature map ϕ_k (associated to the kernel k) by,

$$\phi_k : \mathcal{X} \to \mathcal{H}$$
$$\phi_k(x) := k(x, .).$$

Next, let $\mathcal{H}_k^\circ$ denotes the vector space generated by $\phi(\mathcal{X})$. More formally,

$$\mathcal{H}_k^\circ := \{\sum_{i=1}^m a_i k(., x_i) | a_i \in \mathbb{R}, x_i \in \mathcal{X}, m \in \mathbb{N}\} \subset \mathcal{H}.$$

We can define the inner product on $\mathcal{H}_k^\circ$ first by,

$$\langle k(x, .), k(y, .) \rangle := k(x, y),$$

and then extend above to whole $\mathcal{H}_k^\circ$ by linearity. This makes $\mathcal{H}_k^\circ$ into an inner product space. Notice that above implies that for $f \in \mathcal{H}_k^\circ$,

$$f(x) = \langle f, k(x, .) \rangle.$$

Finally, define $\mathcal{H}_k$ to be the *completion* of $\mathcal{H}_k^\circ$. We refer to $\mathcal{H}_k$ as the *reproducing kernel Hilbert (RHKS)* space associated to the kernel k. Next, we explain how one can naturally map $P(\mathcal{X}, \Sigma)$ into $\mathcal{H}_k$. For $p \in P(\mathcal{X}, \Sigma)$, we can consider the following associated functional on $\mathcal{H}_k$,

$$T_p : \mathcal{H}_k \to \mathbb{R}$$
$$T_p(f) := \mathbb{E}_{x\sim p} f$$

If k satisfies a mild integrality assumption ([[12], Lemma 3]) then T_p would be in $\mathcal{H}_k^*$ (i.e. T_p is a bounded functional) and by *Riesze representation theorem* we can associate an element $\mu_p \in \mathcal{H}_k$ such that $T_p(f) = \langle \mu_p, f \rangle$. Putting all together we explained a mapping,

$$P(\mathcal{X}) \to \mathcal{H}_k$$
$$p \mapsto \mu_p$$

we refer to the above mapping as *mean embedding*. Also, one can show the following explicit form for μ_p,

$$\mu_p(t) = \int_{\mathcal{X}} k(x, t) dp(x).$$

Above leads to the following definition [8] and [25],

Definition 2.1 (Characteristic kernel)**.** *A kernel k is called characteristic if the mean embedding mapping associated to it is injective.*

We assume all kernels that we consider in this note are characteristic unless otherwise noted. Finally, when $\mathcal{X} = \mathbb{R}^d$, we will record the following definitions [29] and [7],

Definition 2.2 (Analytic kernel)**.** *A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called analytic if for any $x \in \mathbb{R}^d$, $k(x, .)$ and $k(., x)$ are analytic functions.*

Definition 2.3 (Translation invariant kernel)**.** *A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called translation invariant if there exist a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that,*

$$k(x, y) = \kappa(x - y).$$

3 Hypothesis testing with RHKS

The material presented in this subsection are mainly based on [12], [11],[13], [14], [7] and [29]. In this section we discuss how one can design hypothesis testing using properties of RHKS. Consider the same setup as in Subsection 2.2. Let k be a given kernel and $\mathcal{H}_k$ be the RKHS associated to k. Let $\mathcal{F}$ denote the unit ball in $\mathcal{H}_k$,

$$\mathcal{F} := \{f \in \mathcal{H}_k | \, \|f\|_{\mathcal{H}_k} \leq 1\}.$$

Define the the *maximum mean discrepancy (MMD)*,

$$MMD(\mathcal{F}, p, q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{x\sim p}(f(x)) - \mathbb{E}_{x\sim q}(f(x))|.$$

Intuitively, we think of MMD as a measure for how good $\mathcal{F}$ distinguishes p from q. We will work with the unit ball $\mathcal{F}$ through rest of this note and for simplicity of notation we $MMD(p, q)$ instead of $MMD(\mathcal{F}, p, q)$. Also, let $f \in \mathcal{F}$ be a function that achieves the supremum in the definition of MMD. We refer to f as *witness function*. Next, It follows that [[8],[25],[12]] when the background kernel k is characteristic then MMD defines a metric on $P(\mathcal{X}, \Sigma)$ which is a key property. Next, we discuss how one can use this notion to develop a hypothesis testing procedure.

3.1 Tests based on MMD

In order to develop a hypothesis testing procedure, we need to estimate MMD from X_m^p and X_n^q. One can consider the following biased estimation of $MMD(\mathcal{F}, p, q)$ as follows,

$$MMD_b(X_m^p, X_n^q) = \sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(x_i^p) - \frac{1}{n} \sum_{i=1}^{n} f(x_i^q)|.$$

Notice that above estimation is biased since there is a supremum term in the definition. The first strategy to obtain statistical test is to understand how MMD_b is concentrated around MMD. This the content of the following theorem,

Theorem 3.1 ([12], Theorem 7 and [11], Theorem 5)**.** *Assume the kernel function satisfies* $0 \leq k(x,y) \leq K$ *for some* $K \in \mathbb{R}$*. Then for* $\epsilon > 0$*, the following is true,*

$$\mathbb{P}(|MMD_b(X_m^p, X_n^q) - MMD(p,q)| \geq 2(\sqrt{\frac{K}{m}} + \sqrt{\frac{K}{n}}) + \epsilon) \leq 2\exp(\frac{-\epsilon^2 mn}{2K(m+n)}),$$

where the probability is over taking m *i.i.d sample from* p *and* n *i.i.d sample from* q*.*

Next, we discuss an application of above theorem. Assume (for simplicity) $m = n$ and the goal is to design a hypothesis testing with level α as we explained in Subsection 2.3. Set the design parameter $\gamma = \sqrt{\frac{2K}{n}}(1 + \sqrt{2\log(\frac{1}{\alpha})})$ and consider the following testing procedure,

$$\tau(X_n^p, X_n^q) = \begin{cases} \text{reject} & MMD_b(X_n^p, X_n^q) \geq \gamma \\ \text{accept} & MMD_b(X_n^p, X_n^q) < \gamma \end{cases}.$$

It follows from above theorem [[12], Corollary 9 and [11], Lemma 6] that τ has level α . However, from practical point of view, there is one more step in-order to have a complete statistical testing procedure. Notice that, given samples X_n^p, X_n^q one needs to compute test statistic $MMD_b(X_n^p, X_n^q)$ but by definition one needs to compute a supremum over the unit ball of $\mathcal{H}_k$ which has infinite elements. How can we compute this in an efficient way? The following proposition answers this question.

Proposition 3.2 ([12], Lemma 6)**.**

$$MMD_b(X_m^p, X_n^q) = \sqrt{\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i^p, x_j^p) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i^p, x_j^q) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i^q, x_j^q)}$$

Proposition 3.2 is useful in the sense that given a kernel function k, one can compute the test statistic $MMD_b(X_m^p, X_n^q)$ directly without testing all functions in the unit ball. Also notice above comes with a quadratic cost $O((m+n)^2)$ for kernel term computation. This completes a statistical testing procedure using concentration bounds. We next explain another strategy to obtain statistical hypothesis testing using by asymptotic distribution of certain test statistic. Next, one can use similar argument that leads to Proposition 3.2 to obtain an *unbiased* test statistics for $MMD(p,q)$ as follows,

Proposition 3.3 ([12], Lemma 6)**.** *Define,*

$$MMD_u(X_m^p, X_n^q) := \sqrt{\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i^p, x_j^p) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i^q, x_j^q) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i^p, x_j^q)}.$$

Then $MMD_u(X_m^p, X_n^q)$ *is unbiased estimator of* $MMD(p,q)$*. Moreover, when* $m = n$ *above can be written in compact form,*

$$MMD_u(X_n^p, X_n^q) := \sqrt{\frac{1}{n(n-1)}\sum_{i\neq j}^{n} H_{ij}},$$

where, $H_{ij} = k(x_i^p, x_j^p) + k(x_i^q, x_j^q) - k(x_i^p, x_j^q) - k(x_j^p, x_i^q)$.

One can study concentrations of $MMD_u((X_m^p, X_n^q))$ around $MMD(p,q)$ and obtain a similar test that we described above. This is the content of [[12], Theorem 10] and [[12], Corollary 11]. One can also study the *asymptotic distribution* of MMD_u to design a hypothesis testing. This is the content of [[12], Theorem 12]. More precisely, under the alternative hypothesis $p \neq q$, the following for *Central Limit* like result is true for the asymptotic distribution of $MMD_u(X_n^p, X_n^q)$ [[19], [12]] ,

Theorem 3.4. *Assume alternative hypothesis* $p \neq q$ *and* $m = n$ *then the following is true,*

$$\sqrt{n}(MMD_u^2(X_n^p, X_n^q) - MMD^2(p,q)) \to \mathcal{N}(0, \sigma^2)$$

where the above convergence is in the sense of distributions and,

$$\sigma^2 = 4\mathbb{E}(H_{12}H_{13}) - 4\mathbb{E}(H_{12}^2)$$

where H_{ij} *are as in Proposition 3.3.*

We will explain how one can generalize and use above result to design hypothesis testing in the following sections. Finally, we also note that one can use similar techniques discussed here to develop a test of independence for random variables [14].

3.2 Faster tests motivated by MMD

In this subsection we consider the same setup as in Subsection 2.2. Moreover, we assume that $\mathcal{X} = \mathbb{R}^d$ and the i.i.d samples $X_n^p := \{x_1^p, ..., x_n^p\}$ and $X_n^q := \{x_1^q, ..., x_n^q\}$ from p and q have equal size n. Motivated by the results stated in previous subsection, the goal here is to design faster tests (e.g linear instead of quadratic). To start, let $k(x, y) = \kappa(x - y)$ be a translation invariant and analytic kernel (see Definition 2.2 and Definition 2.3) on $\mathbb{R}^d$. The following formula is true ([[26], Corollary 4] and [7]),

$$MMD^2(p, q) = \int_{\mathcal{X}} |\phi_p(t) - \phi_q(t)|^2 F^{-1}(\kappa(t))dt$$

where F^{-1} denotes the inverse Fourier transform and ϕ_p and ϕ_q are characteristic functions of p and q respectively. One can try and sample from the measure ν on $\mathcal{X}$ that is defined by $d\nu := F^{-1}(\kappa(t))dt$ to estimate MMD with goal of designing a statistical hypothesis tests but as [7] reports this method does not perform well. However, motivated by above, for a fixed measure ν on $\mathcal{X}$, one can first sample $X_m^\nu := \{x_1^\nu, ..., x_m^\nu\}$ from ν and then consider the following "random distance" between probability measures,

$$d^2_{\mu,m}(p, q) := \frac{1}{m} \sum_{i=1}^{m} |\mu_p(x_i^\nu) - \mu_q(x_i^\nu)|^2.$$

where μ_p, μ_q are the mean embedding of p and q (See Subsection 2.4 for details) and these terms can be estimated by kernel terms as follows,

$$\mu_p(x_i^\nu) \sim \frac{1}{n} \sum_{j=1}^{n} k(x_j^p, x_i^\nu)$$

$$\mu_q(x_i^\nu) \sim \frac{1}{n} \sum_{j=1}^{n} k(x_j^p, x_i^\nu),$$

and we can plug above back into the definition of $d^2_{\mu,m}(p, q)$. More precisely, for $1 \leq i \leq n$,

$$Z_i := (k(x_i^p, x_1^\nu) - k(x_i^q, x_1^\nu), ..., k(x_i^p, x_m^\nu) - k(x_i^q, x_m^\nu)) \in \mathbb{R}^m,$$

and let Z to be the $m \times n$ matrix with Z_i as its columns and set $\Sigma_n := \frac{1}{n} ZZ^T$ to be a $m \times m$ matrix and set $W_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \in \mathbb{R}^m$ and consider the following statistics,

$$S_n := nW_n \Sigma_n^{-1} W_n.$$

It can be proved [[7], Proposition 2] that under the null hypothesis $p = q$, S_n is asymptotically has distribution of a χ^2 random variable with m degrees of freedom. This leads to the following test ,

$$\tau(X_n^p, X_n^q) = \begin{cases} \text{reject} & S_n \geq \gamma \\ \text{accept} & S_n < \gamma \end{cases}.$$

Notice that S_n in the above depends on X_n^p, X_n^q, X_m^ν. However, we can work with a fixed ν and think of the testing procedure depends on X_n^p, X_n^q which matches with our setup in Subsection 2.3. Putting everything together [7] obtains the following testing procedure.

Corollary 3.1 ([7]). *For a given level α if we choose the design parameter γ in the above correspond $1 - \alpha$ quantile of such a χ^2 distribution with m degrees of freedom then the τ will have level α.*

Why above gives a faster test procedure compared to the results in Subsection 3.1? We answer this question in the following. In practice, we can choose m to be small (say 10 as noted in [7]) then the computation of Σ^{-1} would be very fast since Σ is a $m \times m$ matrix. This means that one only needs $O(n)$ kernel computation (to compute Z_i for $1 \leq i \leq n$) which is linear compared to the quadratic test that we explained in Subsection 3.1.

Next, we explain how to generalize even further by discussing recent work [29]. Motivated by previous result we can consider the following as our starting point,

$$d^1_{\mu,m}(p, q) := \frac{1}{m} \sum_{i=1}^{m} |\mu_p(x_i^\nu) - \mu_q(x_i^\nu)|^1.$$

Notice that in the above we are considering the ℓ^1 norm of mean embedding differences instead of ℓ^2 that we did before in [7]. [29] follows along same analysis that is done in [7] but there some key differences that leads to modified statistics and asymptotic distribution. We next explain this with more detail. Let $X_m^\nu := \{x_1^\nu, ..., x_m^\nu\}$ be as before and for $1 \leq i \leq n$, set,

$$Z_i^p := (k(x_i^p, x_1^\nu), ..., k(x_m^p, x_1^\nu)) \in \mathbb{R}^m$$
$$Z_i^q := (k(x_i^q, x_1^\nu), ..., k(x_m^q, x_1^\nu)) \in \mathbb{R}^m$$

Let Σ^p and Σ^q be the $m \times m$ covariance matrix of Z_i^p and Z_i^q respectively and set,

$$\Sigma^{p,q} := 2(\Sigma^p + \Sigma^q),$$

and also set,

$$S^{p,q} := (\mu_p(x_1^\nu) - \mu_q(x_1^\nu), ..., \mu_p(x_1^\nu) - \mu_q(x_1^\nu)) \in \mathbb{R}^m$$

Finally, consider the following statistics,

$$L1(X_n^p, X_n^q)) := \left\| \sqrt{2n} (\Sigma^{p,q})^{-\frac{1}{2}} S^{p,q} \right\|_1,$$

where in the above $\|\|_1$ denotes the ℓ_1 norm in $\mathbb{R}^m$. It is proved [[29], Proposition 3.2] that $L1(X_n^p, X_n^q))$ has the same distribution as sum of m Nakagami random variables [1] with parameters $(\frac{1}{2}, 1)$ which we denote this by $Naka(\frac{1}{2}, 1, m)$. This result can be thought of as an analogue of previous result in [7] that we discussed above where the limiting distribution was χ^2 with m degrees of freedom. Next, one can obtain an analogue of Corollary 3.1 by simply taking the design parameter γ correspond to $1 - \alpha$ quantile of the $Naka(\frac{1}{2}, 1, m)$. It also follows from the same argument we did above that this test is faster than results in Subsection 3.1. The other key advantage of considering $d^1_{\mu,m}$ instead of $d^2_{\mu,m}$ is that the corresponding test has *greater testing power* as it is proved in [[29], Proposition 3.1]. Finally, we briefly mention few other related works. It is possible to consider scores other than MMD for distinguishing p from q. For example one can try to use Stein's identity for which one need to test a class smooth functions to distinguish p from q which makes it computationally difficult. However, in [20] a kernelized approach is developed for this problem which leads to alternative way of hypothesis testing not based on MMD. Finally, [31] studies the two sample problem from a Bayesian point of view.

4 Selecting a kernel

The materials presented in this section are mainly based on [19],[28],[27],[4], [16] and [15]. Consider the same setting as in Subsection 2.2. Assume we want to design the hypothesis for this problem using the techniques that we explained in the previous sections. The main question is how to choose a good kernel for the problem?

4.1 Kernel parametrized by deep nets

Let us assume that we are in a setting that p and q have "complex support" that standard kernels (e.g. gaussian) does not perform well on those. The main idea of [19] is to parametrize a class of kernels by a deep net and optimize the parameter so that the resulting test has a maximum testing power. More specifically, consider,

$$k_w(x, y) := [(1 - \epsilon)\kappa(\phi_w(x), \phi_w(y)) + \epsilon]v(x, y)$$

where $0 < \epsilon < 1$ and κ are standard simple kernels (e.g. Gaussian) ,v is a simple characteristic kernel (Denition 2.1) and $\phi_w : \mathcal{X} \to \mathcal{X}'$ is a neural net which we can think it extracts "complex" features. Why one should consider above parametrization? Recall from Subsection 2.4 and Section 2.3 that one need to work with characteristic kernels and the following shows that above form give a parametrization characteristic kernel.

Proposition 4.1 ([19], Proposition 5)**.** *Let $0 < \epsilon < 1$ and v be a characteristic kernel then*

$$k_w(x, y) := [(1 - \epsilon)\kappa(\phi_w(x), \phi_w(y)) + \epsilon]v(x, y),$$

defines a characteristic kernel.

The next question that we need to address is what is the loss function for the neural net ϕ_w? In [19], this network is trained to maximize the testing power of k_w and needs to estimate the power of k_w using the sample data. In [19], this is done by utilizing the earlier works [27] and [28] which we next explain. For any w we can design a hypothesis testing procedure τ using MMD as we explained in Section 3 . It follows from [27] and [28] that the power of τ is approximately equal to,

$$J(p, q, k_w) := \frac{MMD^2(p, q, k_w)}{\sigma^2}$$

where as we noted in Theorem 3.4,

$$\sigma^2 = 4\mathbb{E}(H_{12}H_{13}) - \mathbb{E}(H_{12}^2).$$

It remains to estimate above using sample data. For the $MMD^2(p, q, k_w)$, we can estimate it using Proposition 3.3 and for the denominator, it follows from [27] and [28] that the following gives a biased estimator (In [19] they a regularizer term as well),

$$\frac{4}{n^3}\sum_{i=1}^{n}(\sum_{j=1}^{n} H_{ij})^2 - \frac{4}{n^4}(\sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij})^2,$$

for σ^2 where H_{ij} are defined in Proposition 3.3. Putting altogether, we can train a neural net ϕ_w to learn a characteristic kernel k_w by maximizing above estimation which implies that the resulting hypothesis testing procedure will have greater testing power. Along this line, [9], uses deep kernel with maximized power to distinguish adversarial data from real data. Finally, we note that [19] can be thought of extension of [15] where the idea of selecting a kernel with maximized testing power is studied as well.

4.2 Other approaches

In this subsection, we briefly talk about other related approaches. In [16], a neural net ϕ_w is trained based on a task constructed from samples X_n^p and X_n^q. Intuitively, we can think that ϕ_w learns important features to distinguish p and q. Then the following kernel is considered for hypothesis testing, $k_w(x, y) := \kappa(\phi_w(x), \phi_w(y))$. As we discussed above this kernel is not necessarily characteristic and there is no provable guarantee in this approach that shows the resulting kernel is optimal in some sense. Finally, [6] studies MMD statistics obtained by special anisotropic kernels which are special class of kernels based on Mahalanobis distance that their covariance matrix is not a multiple identity matrix.

5 Applications

In this section we discuss application of two sample problem and the techniques that we described here in other machine learning problems.

5.1 Applications for GANs

We start by briefly recalling the general setting of Generative Adversarial Networks (GAN) [10] and Wasserstein Generative Adversarial Networks (WGAN) [3] and then explain how these setting can be connected to the techniques that we discussed in this note. Generative Adversarial Networks (GAN) [10] can be thought of as unsupervised learning algorithms to learn a hidden probability measure. More specifically, assume $\mathcal{X}$ is some compact metric space and $p \in P(\mathcal{X}, \Sigma)$ is a hidden probability measure on $\mathcal{X}$. We can think of p corresponding to the distribution of real data that we do not have access to p but we can sample from it. We would like to develop algorithms to learn p using a parametrized family of probability measures p_θ on $\mathcal{X}$ where $\theta \in \Theta$ (we can think of θ as weight of a neural net). More formally, for a good notion of "distance" , d, between probability measures, we can consider optimizing the following,

$$\theta \mapsto d(p_\theta, p)$$

In [10], d in the above is related to b Jensen-Shannon divergence between probability measures and this choice leads to some training issues (e.g vanishing gradients or mode collapse). In [3], this problem is addressed by choosing d to be *Wasserstein distance* and we recall it's definition next [30],[3],

Definition 5.1 (Wasserstein)**.** *For $\mu, \nu \in P(\mathcal{X}, \Sigma)$, the Wasserstein distance (Earth-Mover) is defined by,*

$$W(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\gamma(x, y),$$

where,

$$\Pi(\mu, \nu) := \{\gamma \in P(\mathcal{X} \times \mathcal{X}, \Sigma \times \Sigma) | (\pi_1)_*(\gamma) = \mu, (\pi_2)_*(\gamma) = \nu\},$$

and $c(x, y) = d_{\mathcal{X}}(x, y)$ is a the cost function and π_1 and π_2 are projection on first and second coordinate respectively.

We can think of $W(\mu, \nu)$ as the optimal cost of transporting "probability mass" of ν into μ. It turns out that this notion of distance gives a meaningful loss function especially when the probability measures p and p_θ are supported in a low dimensional sub-manifold as noted in [2] and [3]. However, running gradient descent on,

$$\theta \mapsto W(p_\theta, p)$$

is not obvious because it is not clear from from Definition 5.1 how to differentiate above function. This problem is addressed in [2] by using the following,

Theorem 5.1 (Kantorovich-Rubinstein duality)**.**

$$W(p_\theta, p) = \sup_{Lip(f)\leq 1} \left(\mathbb{E}_{x\sim p}(f(x)) - \mathbb{E}_{x\sim p_\theta} f(x)\right)$$

where, for a function $f : \mathcal{X} \to \mathbb{R}$,

$$Lip(f) := \sup_{x,y\in\mathcal{X}, x\neq y} \frac{f(x) - f(y)}{d_{\mathcal{X}}(x, y)}.$$

Roughly speaking, one needs to run a separate neural net to approximate the function that achieves the supremum in the statement of Kantorovich-Rubinstein duality and use that function for gradient descent. In [2], this function is called "critic". Now if we look at the statement Kantorovich-Rubinstein duality it reminds us $MMD(p_\theta, p)$ with the only difference being that the supremum is taken over 1-Lipschitz function instead of the unit ball in $\mathcal{H}_k$. This motivates to search for "critic" by optimizing $MMD(p, p_\theta)$ over the unit ball. In other words, the "critic" in the WGAN would an analogue of witness function in MMD setting (Section 3 for definition). The advantage of this approach is that we no longer need to train a separate neural net to find the critic because one can estimate it using the kernel(e.g. with similar argument to Proposition 3.3). This approach opens new questions regarding the connection of MMD and GANs. In [4], the authors find an unbiased estimator for gradient updates of critic in MMD-based GAN. In [28], MMD hypothesis testing framework is used to evaluate performance of GANs. Finally, we also note that, one can start with a kernel k and define special metric based on k on $\mathcal{X}$ by, $d(x, y) = \frac{k(x,x)+k(y,y)}{2} - k(x, y)$, then it can be shown [22] that the Wasserstein distance between probability measures using d is related to usual MMD distance. From this point of view, one can also use Wasserstein distance for hypothesis testing and we refer to [22] for more details.

5.2 Other applications

In this subsection we briefly discuss some other applications of the techniques that we discussed in this note in other domains. In [17], the authors studied the problem of identifying applications in a network based on their communication. The communication set of each application with the server is modeled as a probability distribution and MMD distance is used to distinguish application. Next, we briefly discuss an application in *Topological data analysis (TDA)*. Persistence homology groups are an important tool in TDA. Roughly speaking, these groups capture the topological properties in a sampled data (e.g the number of holes). In [23], a kernel is constructed on persistence diagrams and [18] uses this to map probability measures on these diagrams to the corresponding RHKS. As an application one can test the hypothesis that whether two sampled dataset have "same topology" or not. Finally, [21] uses MMD based test for model criticism. In MMD setting one can estimate the witness function using kernel and the key point in [21] is that one can use the witness function to say where the model misrepresents the data.

Bibliography

[1] https://en.wikipedia.org/wiki/Nakagami_distribution.

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[4] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[5] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[6] Xiuyuan Cheng, Alexander Cloninger, and Ronald R Coifman. Two-sample statistics based on anisotropic kernels. *Information and Inference: A Journal of the IMA*, 9(3):677–719, 2020.

[7] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1981–1989. Curran Associates, Inc., 2015.

[8] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20:489–496, 2007.

[9] Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy is aware of adversarial attacks. *arXiv preprint arXiv:2010.11415*, 2020.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513–520, 2006.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[13] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[14] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20:585–592, 2007.

[15] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012.

[16] Matthias Kirchler, Shahryar Khorasani, Marius Kloft, and Christoph Lippert. Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1398. PMLR, 2020.

[17] J. Kohout and T. Pevný. Network traffic fingerprinting based on approximated kernel two-sample test. *IEEE Transactions on Information Forensics and Security*, 13(3):788–801, 2018.

[18] Roland Kwitt, Stefan Huber, Marc Niethammer, Weili Lin, and Ulrich Bauer. Statistical topological data analysis - a kernel perspective. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3070–3078. Curran Associates, Inc., 2015.

[19] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Dougal J Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020.

[20] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284, 2016.

[21] James R Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837, 2015.

[22] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[23] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

[24] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[25] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory (COLT 2008)*, pages 111–122. Omnipress, 2008.

[26] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[27] Dougal J. Sutherland. Unbiased estimators for the variance of mmd estimators. *https://arxiv.org/abs/1906.02104*, 2019.

[28] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

[29] Gael Varoquaux et al. Comparing distributions: ℓ_1 geometry improves kernel two-sample testing. In *Advances in Neural Information Processing Systems*, pages 12327–12337, 2019.

[30] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[31] Qinyi Zhang, Sarah Filippi, Seth Flaxman, and Dino Sejdinovic. Bayesian kernel two-sample testing. *arXiv preprint arXiv:2002.05550*, 2020.