On the High Dimensional Power of a Linear-Time Two Sample Test under Mean-shift Alternatives

Sashank J. Reddi[*]
sjakkamr@cs.cmu.edu

Aaditya Ramdas[*]
aramdas@cs.cmu.edu

Barnabás Póczos
bapoczos@cs.cmu.edu

Aarti Singh
aarti@cs.cmu.edu

Larry Wasserman
larry@stat.cmu.edu

Carnegie Mellon University, Pittsburgh, USA

Abstract

Nonparametric two sample testing deals with the question of consistently deciding if two distributions are different, given samples from both, without making any parametric assumptions about the form of the distributions. The current literature is split into two kinds of tests - those which are consistent without any assumptions about how the distributions may differ (*general* alternatives), and those which are designed to specifically test easier alternatives, like a difference in means (*mean-shift* alternatives). The main contribution of this paper is to explicitly characterize the power of a popular nonparametric two sample test, designed for general alternatives, under a mean-shift alternative in the high-dimensional setting. Specifically, we explicitly derive the power of the linear-time Maximum Mean Discrepancy statistic using the Gaussian kernel, where the dimension and sample size can both tend to infinity at any rate, and the two distributions differ in their means. As a corollary, we find that if the signal-to-noise ratio is held constant, then the test's power goes to one if the number of samples increases faster than the dimension increases. This is the first explicit power derivation for a general nonparametric test in the high-dimensional setting, and the first analysis of how tests designed for general alternatives perform against easier ones.

1 Introduction

The central topic of this paper is nonparametric two-sample testing, in which we try to detect a difference between two d-dimensional distributions P and Q based on n samples from both, i.e. deciding whether two samples are drawn from the same distribution. We will be concerned with the following two settings, the first of which deals with *general* alternatives (GA), i.e.

$$H_0 : P = Q \quad \text{v.s.} \quad H_1 : P \neq Q. \tag{GA}$$

It is called nonparametric two-sample testing because no parametric assumptions are made about the form of P, Q (like Gaussianity or exponential families). We use the term *general* alternatives to mean that the difference between P, Q need not have a simple form. In contrast, the second setting that we are concerned about deals with *mean-shift* alternatives (MSA), i.e.

$$H_0 : \mu_P = \mu_Q \quad \text{v.s.} \quad H_1 : \mu_P \neq \mu_Q \tag{MSA}$$

where $\mu_P = \mathbb{E}_{X \sim P}[X]$ and $\mu_Q = \mathbb{E}_{Y \sim Q}[Y]$. It is still nonparametric two-sample testing, since we make no assumptions about P, Q, but deals with *easier* alternatives, meaning that we specify the exact form in which P and Q differ, i.e. they differ in their means. Parametric two-sample testing (for example, when P, Q are Gaussian) is also important, but will be out of the scope of our discussion; see Lopes et al. (2011) for a recent example. We assume equal number n of samples for simplicity; our results would also go through if $n_1/(n_1 + n_2) \to c \in (0, 1)$ as $n_1, n_2 \to \infty$.

1.1 Hypothesis testing terminology

Let $X^{(n)} = \{x_1, ..., x_n\} \sim P$ and $Y^{(n)} = \{y_1, ..., y_n\} \sim Q$ be the two sets of samples, where $x_i, y_j \sim \mathbb{R}^d$ for all $1 \leq i, j \leq n$. A *test* is any function or algorithm that

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

takes $X^{(n)}, Y^{(n)}$ as input, and outputs $\{0, 1\}$ where 1 is interpreted to mean that it *rejects the null hypothesis* H_0, and 0 is interpreted to mean that *there is insufficient evidence to reject* H_0. A test is characterized by its false positive rate or type-1 error

$$\alpha = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$$

and its false negative rate or type-2 error

$$\beta = P(\text{not rejecting } H_0 \mid H_1 \text{ is true}).$$

There is usually a tradeoff involved - decreasing one error rate increases the other. Hence, one sometimes fixes α to some small value (say 0.01), and refers to $\phi = 1 - \beta$ as the *power of the test at* $\alpha = 0.01$. A test is classically called consistent if for any fixed α, the power $\phi \to 1$ as $n \to \infty$ whenever H_0 is false.

Many tests in the literature, including the ones we will consider, calculate a test statistic T (as a function of $X^{(n)}, Y^{(n)}$), and reject the null hypothesis if $T > c_\alpha$, where the threshold c_α depends on the distribution of T under H_0 and on a pre-defined α. See Lehmann & Romano (2006) for a detailed introduction.

1.2 Motivation

Our first motivation comes from the fact that there is a big difference between the classical setting of fixing d while letting $n \to \infty$, and the high-dimensional (HD) setting obtained when

$$(n, d) \to \infty \qquad \text{(HD)}$$

A test would be called consistent under HD if for any fixed α, the power $\phi \to 1$ as $(n, d) \to \infty$ whenever H_0 is false. It is of vital importance, both theoretically and practically, to understand the power of tests in such settings, and to characterize the rate at which n must grow as a function of d so that the test is still consistent. While classical tests were proposed for the low-dimensional settings, over the past two decades several tests have been proposed specifically for MSA and studied in the HD setting; see Subsection 1.3. However, to the best of our knowledge there has been no formal and precise characterization of power of tests designed for GA in high dimensions.

Our second motivation comes from the observation that there is no literature on how tests designed for GA perform under MSA. In other words, while it is expected that tests designed for MSA will not be consistent against more general GA, it is unclear how exactly tests designed for general alternatives fare when when faced with a mean-shift alternative.

1.3 Related Work (MSA)

It is well known (see Kariya (1981); Simaika (1941); Anderson (1958); Salaevskii (1971)) that if P, Q are Gaussians, then the uniformly most powerful test in the fixed-dimension setting under fairly general conditions, is the T-test by Hotelling (1931) :

$$T_H := (m_P - m_Q)^T S^{-1} (m_P - m_Q)$$

where m_P, m_Q and S are the usual empirical estimators of μ_P, μ_Q and the joint covariance matrix Σ. In a seminal paper, Bai & Saranadasa (1996), showed that in the high-dimensional setting, the T-test performs quite poorly (specifically when $(n, d) \to \infty$ with $d/n \to 1 - \epsilon$ for small ϵ). This is intuitively because of the difficulty of estimating the $O(d^2)$ parameters of Σ^{-1} with very few samples. Indeed, S^{-1} is not even defined when $d > n$ and is poorly conditioned when d is of similar order as n. To avoid this problem, they proposed to use the test statistic

$$T_{BS} := (m_P - m_Q)^2 - \mathrm{tr}(S)/n$$

T_{BS} has non-trivial power when $d/n \to c \in (0, \infty)$. Srivastava & Du (2008) proposed to instead use $\mathrm{diag}(S)$ instead of S in T_H, and showed its advantages in certain settings over T_{BS}. More recently, Chen & Qin (2010), henceforth called CQ, proposed a slight variant of T_{BS}, which is a U-statistic of the form

$$T_{CQ} := \frac{1}{n(n-1)} \sum_{i \neq j}^{n} (x_i^T x_j + y_i^T y_j) - \frac{2}{n^2} \sum_{i,j=1}^{n} x_i^T y_j$$

that achieves the same power without explicit restrictions on d, n, but rather in terms of conditions stated in terms of $n, \mathrm{tr}(\Sigma), \mu_P - \mu_Q$. The settings of under which these various statistics are consistent, or achieve non-trivial power, are slightly complicated to describe, and the reader is referred to their papers for details.

1.4 Related Work (GA)

There are many nonparametric test statistics for two-sample testing. One of the most popular tests is the kernel Maximum Mean Discrepancy, henceforth called MMD, proposed in Gretton et al. (2012). While the technical details of the kernel literature are unnecessary for the purposes of this paper, it suffices to say that the population statistic is

$$\mathrm{MMD} := \max_{\|f\|_H \leq 1} \mathbb{E}_P f(x) - \mathbb{E}_Q f(y)$$

where H is a Reproducing Kernel Hilbert Space and $\|f\|_H \leq 1$ is its unit norm ball. There are two related sample statistics, both of which can be shown to be unbiased estimators of MMD. The first is a U-statistic

$$\begin{aligned}\mathrm{MMD}_u^2 &= \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(x_i,x_j)\\ &+ \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(y_i,y_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i,y_j)\end{aligned}$$

The second is a linear-time statistic

$$\begin{aligned}\mathrm{MMD}_l^2 &= \frac{1}{n/2}\sum_{i=1}^{n/2}[k(x_{2i-1},x_{2i}) + k(y_{2i-1},y_{2i})\\ &\quad -k(x_{2i-1},y_{2i}) - k(y_{2i-1},x_{2i})]\end{aligned}$$

Note that T_{CQ} is just MMD_u^2 under the linear kernel $k(x,y)=x^Ty$. It is known that in the fixed d setting, the power of both MMD_l^2 and MMD_u^2 approaches 1 at the rate of $\Phi(\sqrt{n})$ where Φ is the standard normal cdf, see Gretton et al. (2012). However, nothing is formally known when d could be increasing with n.

A recent related manuscript by Ramdas et al. (2015) conducts detailed experiments that demonstrate that in the fixed n, increasing d setting, the power of MMD and distance correlation decay *polynomially* in high dimensions against fair alternatives. While the authors provide some initial insights into this phenomenon for specific examples, there is still no theoretical analysis of the power of MMD (or any statistic designed for GA) against MSA or GA or any other set of alternatives, in the high dimensional setting.

Another statistic called Energy Distance by Székely & Rizzo (2004) is closely tied to the MMD - indeed it has the same form as the MMD with the Euclidean distance instead of a kernel; Lyons (2013) showed that one can also use other metrics instead of the Euclidean distance and Sejdinovic et al. (2013) showed that there is a close tie between metrics and kernels for these problems. There has been an initial attempt to characterize some properties of distance correlation (which is a related statistic for the related problem of independence testing) in high dimensions in Székely & Rizzo (2013), but no analysis of power is available or easily derivable. There also exist many other tests under GA like the cross-match test by Rosenbaum (2005), but none of them have been analyzed under HD.

2 Power of MMD_l (fixed dimension)

Let us first review the basic argument from Gretton et al. (2012) showing the power in the fixed dimensional setting. It will then become clear what the main difficulties are in establishing results in the high-dimensional setting.

The main tool needed is a simple convergence result of the sample statistic to the population quantity. It becomes convenient to introduce the notation $z_i=(x_i,y_i)$ and $h_{ij}=h(z_i,z_j)$ where

$$h_{ij} := k(x_i,x_j)+k(y_i,y_j)-k(x_i,y_j)-k(x_j,y_i). \qquad (1)$$

Then we can rewrite our test statistic as

$$\mathrm{MMD}_l^2 = \frac{1}{n/2}\sum_{i=1}^{n/2} h(z_{2i-1},z_{2i}). \qquad (2)$$

Its expectation is $E_{z,z'}h(z,z') = \mathrm{MMD}^2$ and then Corollary 16 of Gretton et al. (2012) states that under both H_0 and H_1, we have

$$F := \frac{\sqrt{n}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} \rightsquigarrow N(0,1) \qquad (3)$$

where $V = 2\mathrm{Var}_{z,z'}h(z,z')$ and $\rightsquigarrow$ means convergence in distribution as $n\to\infty$. Note that V is a constant independent of n, and so there exists a constant z_α such that $P(Z>z_\alpha)\le\alpha$ when $Z\sim N(0,1)$. Then, the corresponding test rejects H_0 whenever

$$\text{Test-MMD}_l^2 \quad : \quad \frac{\sqrt{n}\mathrm{MMD}_l^2}{\sqrt{v}} > z_\alpha \qquad (4)$$

where v is twice the empirical variance of $h(z,z')$. If Pr denotes the probability under H_1, the power of this test is given by

$$\begin{aligned} & \Pr\left(\frac{\sqrt{n}\mathrm{MMD}_l^2}{\sqrt{v}} > z_\alpha\right) & (5)\\ = \; & \Pr\left(F > \sqrt{\frac{v}{V}}z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) & (6)\\ \xrightarrow{n\to\infty} \; & \Pr\left(Z > z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) & (7)\\ = \; & 1-\Phi\left(z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) & (8)\\ = \; & \Phi\left(\frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}} - z_\alpha\right) & (9)\end{aligned}$$

where Φ is the standard normal cdf. This behaves like $\Phi(\sqrt{n})$ since the population MMD^2 and V are constants that are both independent of n.

2.1 The challenges in high dimensions

There are several significant difficulties in lifting this argument to the high-dimensional setting.

C1. The population MMD depends on dimension (via the signal strength and bandwidth, as we later show), and one needs to explicitly account for this.

C2. The variance V also depends on dimension (and the signal strength, and the bandwidth, as we later show), and again one needs to explicitly track this, especially its dependence on dimension.

C3. In the increasing d, n setting, the limiting distribution is no longer trivially normal, and one needs to establish conditions under which it is indeed normal - the most important question being if the rate of convergence to normality depends on d.

C4. In the increasing d, n setting, one needs to characterize the rate at which v/V still tends to 1, so that $\sqrt{\frac{v}{V}} z_\alpha$ converges to z_α - since v, V depend on d, the key question is again whether the rate of convergence depends on d or not.

We will have to account for each of these challenges explicitly, as we shall see in later sections. Let us first summarize and discuss our assumptions and contributions before we delve into the technical details.

3 Assumptions and Contributions

We are now in a position to clearly state our contributions. We focus on analyzing the power of MMD_l in the high-dimensional setting when $(n, d) \to \infty$ for the Gaussian kernel with bandwidth γ, i.e. $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{\gamma^2}\right)$, in the mean-shift setting when P and Q differ in their means. Let us first outline our assumptions below; note that we comment about these assumptions in the next subsection.

A1. $x_i = Us_i + \mu_P$ and $y_i = Ut_i + \mu_Q$, where, s_i, t_i are i.i.d random vectors for $i \in \{1, ..., n\}$, each having d i.i.d. zero-mean coordinates.and U corresponds to a $d \times d$ orthogonal rotation i.e. $UU^T = I$.

A2. The k-th central moments of each (i.i.d.) coordinate of s, t exist for $2 \le k \le 6$.

Note that the coordinates of x, y need not be independent and $\mathbb{E}_{x\sim P}[X] = \mu_P, \mathbb{E}_{y\sim Q}[Y] = \mu_Q$. Denote $\delta := \mu_P - \mu_Q$. Denote the second, third and fourth central moments of each i.i.d. coordinate of s, t by σ^2, μ_3, μ_4. Remember that $\mathbb{E}h(z_i, z_j) = \mathrm{MMD}^2$ (see Eq.(2)). Denote the second, third and fourth central moments of $h(z_i, z_j)$ by V, τ_3, τ_4. Let $\|.\|$ represent the Euclidean norm. Our main contribution is:

Theorem 1. *For the Gaussian kernel with bandwidth chosen as $\gamma = \Omega(\sqrt{d})$, under assumptions A1, A2, with $(n, d) \to \infty$ at any rate, the Test-MMD_l^2 (Eq. 4) has asymptotic type-1 error α and asymptotic power*

$$\beta = \Phi\left(\frac{\sqrt{n}\ \|\delta\|^2}{\sqrt{8d\sigma^4 + 8\sigma^2\|\delta\|^2}} - z_\alpha\right)$$

where Φ is the cdf of a standard Normal distribution and z_α is the $(1-\alpha)$ quantile of the standard Normal distribution. For finite samples, type-1 error behaves like $\alpha + 20/\sqrt{n}$ and the power like $\beta - 20/\sqrt{n}$.

The first remarkable point about this theorem is that the power is independent of bandwidth γ, as long as $\gamma = \Omega(\sqrt{d})$. Such behavior has already been noted (but not explained) in the experiments of Ramdas et al. (2015) and we will verify this carefully in our experiments section. While this may not hold true for other kernels, like the Laplace kernel $k(x, y) = \exp\left(-\frac{\|x-y\|_1}{\gamma}\right)$, or against more general alternatives, it is both surprising and interesting that this is the case for the Gaussian kernel under MSA. As discussed later, this theorem applies to the bandwidth chosen by the so-called *median heuristic*; see Schölkopf & Smola (2002). It implies that the median heuristic provides an arguably safe choice in the light of having no further information, and also why it works reasonably well in practice/simulations.

If we consider the *signal to noise ratio* (henceforth called SNR) to be defined as $\Psi := \|\delta\|/\sigma$, then focusing on the more important first term, the power behaves like

$$\Phi\left(\frac{\sqrt{n}\ \Psi^2}{\sqrt{8d + 8\Psi^2}} - z_\alpha\right).$$

From this, we get the following two corollaries. The first applies to the small SNR regime (which includes the *fair* alternative setting, see Ramdas et al. (2015) for details), and the second applies when SNR is large.

Corollary 1. *When the signal to noise ratio Ψ is small, specifically $\Psi = o(d^{1/2})$, the power goes to 1 at the rate of $\Phi(\sqrt{n}\Psi^2/\sqrt{d})$.*

Corollary 2. *When the signal to noise ratio Ψ is large, specifically $\Psi = \omega(d^{1/2})$, then the power goes to 1 at the rate of $\Phi(\sqrt{n}\Psi)$, independent of d.*

Note that the switch in behavior between the two corollaries occurs at Ψ being on the order of $d^{1/2}$, and at this point the prediction of the two corollaries match - hence one could use O, Ω instead of o, ω for describing growth of Ψ in both corollaries.

3.1 Remarks about assumptions

Assumptions (A1,A2) are general enough for the predictions made by our theorem to be accurate and representative of observed behavior. We will verify the predictions of the theorem, corollaries (and later lemmas) in our simulations.

A1. While the coordinates of x, y need not be independent, the first assumption does restrict their covariances to be $\sigma^2 I$. We note that Székely & Rizzo (2013) makes a more restrictive assumption of independent coordinates, while Assumption (a) in Bai & Saranadasa (1996) and Eq.(3.1) in Chen & Qin (2010) assume the same model as we do but don't require spherical covariance. However, our assumption is truly only for mathematical convenience; if we instead had $UD^{1/2}$ in A1, where D is a diagonal rescaling, all our calculations can still be carried out, but would be more tedious since the coordinates of $D^{1/2}s$ are still independent but *not* identically distributed, and we would need to track $\sigma_j^2, \mu_{3j}, \mu_{4j}$ in Appendix Sections 3-6.

A2. The existence of third and fourth moments is needed for calculating population MMD and variance terms, as well as for the Berry-Esseen lemma to control the deviation from normality, and the convergence of v to V. The existence of the sixth moment is needed to bound the Taylor expansion residual term in all our calculations. Note that CQ needs the existence of eighth moments, and BS assume the existence of fourth moments (see Eq. (3.2) in Chen & Qin (2010)) and Assumption (a) in Bai & Saranadasa (1996).

3.2 Remark about bandwidth choice

Remember that the power is independent of the bandwidth γ, as long as $\gamma = \Omega(\sqrt{d})$. This restriction of $\gamma = \Omega(\sqrt{d})$ is to allow us to control the residual term in the Taylor expansion of the Gaussian kernel. However, it is not very restrictive, since smaller γ typically leads to worse power. Specifically, we note that the experiments in Ramdas et al. (2015) for mean-shift alternatives show convincingly that when γ is chosen to be a constant or d^α for $\alpha < 0.5$ (including constant γ), then the power of MMD is poor, while when the highest power occurs for values $\alpha \geq 0.5$. Hence our choice covers most reasonable choices of bandwidth. Furthermore, one of the most popular methods for bandwidth selection is called the *median heuristic*, see Schölkopf & Smola (2002), where one chooses the bandwidth as the median of distances between all pairs of points. A simple calculation shows $\mathbb{E}_{x\sim P, y\sim Q}\|x-y\|^2 = 2\sigma^2 d + \|\mu_P - \mu_Q\|^2$, so generally speaking the median heuristic chooses γ of the same order as $\sigma\sqrt{2d}$ (or larger if $\|\mu_P - \mu_Q\|$ is large).

3.3 Comparisons to CQ

The assumptions in CQ, BS, SD are slightly differently stated from our results here. However, their results can broadly be compared to ours. We can summarize the most recent results, those of CQ, under (A1) and (A2) in the following two observations.

The first observation follows from Eq. (3.11) in Chen & Qin (2010) which applies to the small SNR regime dictated by Eq. (3.4).

Observation 1. *When the signal to noise ratio Ψ is small, specifically $\Psi = o(\sqrt{d/n})$, the power goes to 1 at the rate of $\Phi(n\Psi^2/\sqrt{d})$.*

We believe there is a mistake in the derivation of Eq. (3.12) in Chen & Qin (2010) which applies in the small SNR regime dictated by Eq. (3.5). We describe this in more detail in the Appendix Section 1, and just summarize the corrected resulting observation below.

Observation 2. *When the signal to noise ratio Ψ is large, specifically $\Psi = \omega(\sqrt{d/n})$, then the power goes to 1 at the rate of $\Phi(\sqrt{n}\Psi)$, independent of d.*

Comparing these expressions with Corollary 1 and 2, it is clear that CQ has an advantage over MMD_l in the low-SNR setting. For example, when $n = d$ and the SNR Ψ is constant, the power of CQ can increase $\sqrt{n}$ times faster than that of MMD_l but when the SNR is $\omega(d^{1/2})$, the power of both methods scales in the same fashion. This advantage for low SNR might be wiped out by considering MMD_u^2 - ascertaining if this is the case is an important direction of future work. The main technical challenge is understanding the limiting distributions of general degenerate U-statistics in high dimensions (which in fixed dimensional setting is an infinite sum of χ^2s; see Serfling (2009), Section 5.5.2).

We now provide the proof of Theorem 1 and then verify all our claims in simulations, to convincingly show that these expressions are tight up to constant factors.

4 Proof of Theorem 1

We split the proof into four subsections, one for each of the challenges (C1)-(C4). For C1 and C2, we need to calculate the first two moments of h, introduced in Eq.(1), for which the main tool we use is Taylor expansions (whose validity is explained in Appendix Section 2), following which the results follow after a sequence of tedious calculations and detailed book-keeping. For C3 and C4, we need to bound the third and fourth moments of h. The main tool used for C3 is a Berry-Esseen theorem which helps us track the deviation from normality at finite samples, and C4 is tackled by Chebyshev's inequality once we have a handle on the variance of v. Most of the details will be deferred to the Appendix, but we will outline the main steps of the derivations here.

4.1 The Population MMD

The main takeaway point of the following lemma is the dependence of population MMD^2 on the bandwidth γ and the signal strength $\|\delta\|$ (recall $\delta := \mu_P - \mu_Q$). If p, q are the pdfs of P, Q, then note that the population MMD^2 with the Gaussian kernel is given by

$$\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{\gamma^2}} (p(x)p(y) + q(x)q(y) - 2p(x)q(y))dxdy$$

Lemma 1. *Under (A1),(A2), and when $\gamma = \Omega(\sqrt{d})$ we have*

$$\mathrm{MMD}^2 = \frac{2\|\delta\|^2}{\gamma^2}(1 + o(1)).$$

Proof. We defer details to the Appendix Section 3. On using Taylor's expansion for the Gaussian kernel, the terms in the aforementioned MMD^2 expression can be approximated by bounding higher order residual terms. We prove that the first MMD^2 term is

$$\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{\gamma^2}} p(x)p(y)dxdy = \left(1 - \frac{2\sigma^2}{\gamma^2}\right)^d.$$

Using similar techniques we can also deduce:

$$\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{\gamma^2}} p(x)q(y)dxdy = \prod_i \left(1 - \frac{2\sigma^2}{\gamma^2} - \frac{\delta_i^2}{\gamma^2}\right).$$

Combining these, again using Taylor expansions, gives us our expression. □

4.2 The Variance

As argued earlier, the variance is given by $2V/n$ where $V = \mathrm{Var}_{z,z'} h(z, z')$. The takeaway points of the following lemma are the identical dependence that $\sqrt{V}$ has on bandwidth γ as the MMD^2 (which then causes their ratio to be essentially independent of γ), and also the role played by dimension and the signal strength in determining the variance.

Lemma 2. *Under (A1),(A2), and when $\gamma = \Omega(\sqrt{d})$, we have*

$$V = \frac{16d\sigma^4 + 16\sigma^2\|\delta\|^2}{\gamma^4}(1 + o(1)).$$

Proof. Note that $V = \mathbb{E}_{z,z'} h^2(z, z') - \mathrm{MMD}^4$ since $\mathrm{MMD}^2 = \mathbb{E}_{z,z'} h(z, z')$. Let us focus on the first term:

$$\begin{aligned}\mathbb{E}_{z,z'}[h^2(z, z')] &= \mathbb{E}_{x,x' \sim P} k^2(x, x') + \mathbb{E}_{y,y' \sim Q} k^2(y, y') \\ &+ 2\mathbb{E}_{x \sim P, y \sim Q} k^2(x, y) \\ &+ 2\mathbb{E}_{x,x' \sim P, y,y' \sim Q} k(x, x')k(y, y') \\ &+ 2\mathbb{E}_{x,x' \sim P, y,y' \sim Q} k(x, y')k(x', y) \\ &- 4\mathbb{E}_{x,x' \sim P, y \sim Q} k(x, x')k(x, y) \\ &- 4\mathbb{E}_{x \sim P, y,y' \sim Q} k(x, y)k(y, y')\end{aligned}$$

Hence, there are five different kinds of terms to calculate (the first and last two are similar). Combining these gives us our solution. The details are tedious and hence are given in the Appendix Section 4. □

4.3 The Berry-Esseen Bound

Lemma 3. *Under (A1), (A2), and when $\gamma = \Omega(\sqrt{d})$, we have*

$$\sup_t \left| \mathbb{P}\left(\frac{\sqrt{n/2}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} \le t \right) - \Phi(t) \right| \le \frac{20}{\sqrt{n}}$$

Proof. The Berry-Esseen Lemma (see for example Theorem 3.6 or 3.7 in Chen et al. (2010)), when translated to our problem, essentially yields the above lemma, except that the right hand side is

$$10 \frac{\xi_3}{V^{3/2}\sqrt{n}} \tag{10}$$

where $\xi_3 = \mathbb{E}[|h(z, z') - \mathbb{E}h(z, z')|^3]$, and the constant 10 is not optimal. Note that $\xi_3 \neq \tau_3$ (third central moment of h) due to the absolute value sign. Given that we have the mean and second central moment of h (MMD^2 and V respectively), one might imagine using similar techniques to calculate ξ_3. However, the absolute value poses a problem, and so we must take an alternate route. Specifically, tedious calculations in the Appendix Section 5 prove that τ_4 (the fourth central moment of h) is bounded as

$$\tau_4 \le (4 + o(1))V^2,$$

allowing us to bound ξ_3 as

$$\xi_3 \le \sqrt{\tau_4}\sqrt{V} \le 2V^{3/2}$$

since $\mathbb{E}|X|^3 \le \sqrt{\mathbb{E}|X|^4}\sqrt{\mathbb{E}|X|^2}$ by Cauchy-Schwarz. Substituting into Eq.(10) gives us our Lemma. □

The main challenge involved is in proving that the ratio $\xi_3/V^{3/2}$ is independent of d. Note that a very crude bound of $|h - \mathbb{E}h| \le 4$ (since $e^{-z} \le 1$) gives us $\xi_3 \le 4V$, which would yield a dimension dependence due to an extra $\sqrt{V}$ factor, but because τ_4 (and hence ξ_3) has exactly the right scaling with V, the dependence on V (and hence, importantly, the dimension) cancels out and our Lemma follows. This is only one of the reasons we needed a bound on τ_4, the other appearing in the next lemma.

4.4 Bounding $\sqrt{v/V}$

Recall that v is the empirical estimator of V - it is an empirical average of $n/2$ unidimensional terms. The subtlety is that v depends on d since V depends on d.

What matters is whether the rate of convergence of their ratio to 1 depends on d - fortunately it does not.

Lemma 4. *Under (A1),(A2), and when $\gamma = \Omega(\sqrt{d})$, we have*

$$\sqrt{v/V} = 1 + O_P(1/n^{1/4})$$

Proof. Using $k = 2$ in Theorem A of Section 2.2.3 in Serfling (2009), the bias of v is given by

$$\mathbb{E}[v] - V = -\frac{2V}{n}$$

and its variance is given by

$$\text{var}(v) = \frac{\tau_4 - V^2}{n} \leq \frac{3V^2}{n}$$

both up to smaller order terms (where the inequality follows from the previous lemma).

Then, it is easy to see that $v = V\left(1 + O_P\left(\frac{1}{\sqrt{n}}\right)\right)$, i.e. $v - V = O_P(V/\sqrt{n})$. This is because for any $\epsilon > 0$,

$$\begin{aligned} & P\left(\left|\frac{v-V}{V/\sqrt{n}}\right| > \frac{3+2\sqrt{\epsilon}}{\sqrt{\epsilon}}\right) \\ = \; & P\left(|v - \mathbb{E}[v]| > \frac{3V + 2V\sqrt{\epsilon}}{\sqrt{n\epsilon}} - \frac{2V}{n}\right) \\ \leq \; & \frac{\text{var}(v)}{\left(\frac{3V}{\sqrt{n\epsilon}} + \frac{2V}{\sqrt{n}} - \frac{2V}{n}\right)^2} \\ \leq \; & \epsilon \end{aligned}$$

where we used Chebyshev's inequality, and the second inequality follows since $\frac{3V}{\sqrt{n\epsilon}} + \frac{2V}{\sqrt{n}} - \frac{2V}{n} \geq \frac{3V}{\sqrt{n\epsilon}}$.

□

At this point we have all the key elements of the proof of Theorem 1. Specifically, equations (5) to (9) follow exactly as written, with the exception of (7) holding even with a $\xrightarrow{n,d\to\infty}$ - note that this step allows n, d to grow at any relative rate to ∞ precisely because the rate at which Q converges to the standard normal Z (Berry-Esseen bound) and the rate at which v/V converges to 1, were both independent of d and only needs $n \to \infty$. The dependence on d only enters through the MMD^2 and its variance.

This concludes the proof of Theorem 1. One can also write down the finite sample type-1 error rate as being at most $\alpha + 20/\sqrt{n}$ and the finite sample power as being at least $\beta - 20/\sqrt{n}$, where the additional error is introduced due to the Berry-Esseen bound (whose constants we don't optimize, but could be tightened to about 15 instead of 20).

We now confirm the tightness of all the predictions in this section by detailed simulations in the next section.

5 Experiments

Our aim in this section is to confirm the theoretical predictions made by our lemmas and theorems. The most important claims to address are that the Berry-Esseen bound is independent of d, the null and alternate distributions are indeed normal even in the extreme case when n is fixed and d is increasing, the ratio of $\text{MMD}^2/\sqrt{V}$ is (essentially) independent of the bandwidth, and finally the final power expression is (essentially) independent of the bandwidth and has the exact predicted scaling as given by our expressions.

5.1 Berry-Esseen bound is independent of d

Since the calculations of τ_4 are rather tedious, let us also verify the prediction made in Subsection 4.3 that $\xi_3/V^{3/2}$ is constant and independent of dimension (remember that the ratio involves population quantities). To verify this, we draw 1000 samples from P, Q, and calculate the empirical ratio for d ranging from 40 to 1000, in steps of 20. We make 3 sets of choices for P, Q - standard normals with $\gamma = d^{0.75}$, t_4 distribution with $\gamma = d^{0.5}$ and t_4 distribution with $\gamma = d$. The reason we use t_4 (t distribution with 4 degrees of freedom) is because it does not have a finite fourth moment τ_4. We find that in all 3 cases, the ratio is a constant of about 1.65, showing that our prediction is extremely accurate. Also, while our proof proceeded via bounding τ_4, it seems to hold true even when higher moments than 3 don't exist, since it holds for the t_4 distribution. The spikes are because we calculate a single empirical ratio at each d.

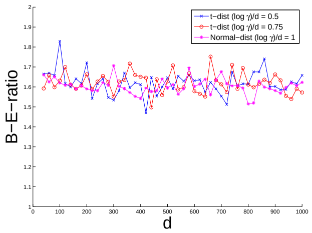

Figure 1: The empirical Berry-Esseen ratio $\xi_3/V^{3/2}$ vs dimension, when $n = 1000$ for the distributions t_4, t_4 and normal, with bandwidths $d^{0.5}, d, d^{0.75}$ respectively.

5.2 Normality of null/alternate distributions

Let us now verify that the null and alternate distributions are indeed (almost) standard normal when n is held constant and d is increased. We do this by fixing $n = 50$, and choosing $d \in \{50, 100, 200\}$ and calculating our test statistic $\sqrt{n}\text{MMD}_l^2/\sqrt{v}$. We experimentally approximate the null and alternate distributions by repeating this process 1000 times; the histogram obtained is compared to a normal by plotting a standard normal quantile-quantile plot. The overlapping straight lines indicate that each of the null and alternate distributions (for three different d values) are almost exactly standard normal even at a small value of n like 50. This agrees with our derivation that the Berry-Esseen constant is very small and normality is achieved soon.

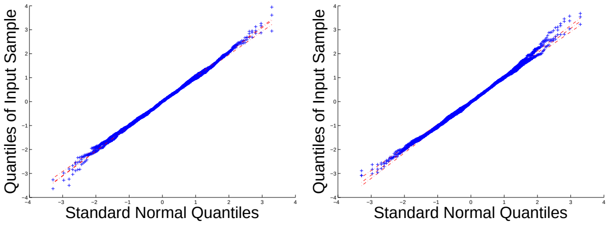

Figure 2: A normal quantile-quantile plot of null (left) and alternate (right) distributions of our test statistic for $d = 50, 100, 200$ when $n = 100$ (1000 repetitions).

5.3 $\mathrm{MMD}^2/\sqrt{V}$ is independent of bandwidth

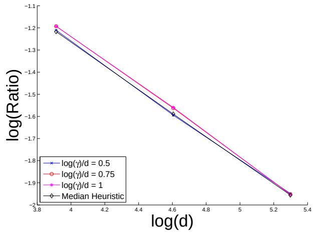

Figure 3: A log-log plot of $\mathrm{MMD}^2/\sqrt{V}$ vs dimension for different bandwidth choices when $\Psi = 1$ and n is large. Note that the slope is -0.5, independent of γ.

Our first two lemmas together imply that the ratio $\mathrm{MMD}^2/\sqrt{V}$ is independent of γ as long as $\gamma = \Omega(\sqrt{d})$. To test this, we actually calculate this ratio for $\gamma = d^{0.5}, d^{0.75}, d$. Remember that these are population quantities - we will estimate the ratio using sample quantities using a large n, when $\Psi = 1$. We plot the obtained log-ratio against log-dimension in Figure 3, showing that the power scales as $1/\sqrt{d}$ as predicted.

5.4 The scaling of power with n, d

Here are a few testable predictions of Theorem 1:

1. When $n = 50$ and $\Psi = 2.5$, the power should decrease as $1/\sqrt{d}$ (Corollary 1).
2. When $n = 50$, and $\Psi = d^{1/4}$, then the power should be a constant (Corollary 1).
3. When $n = d$, and $\Psi = 2$, the power should stay constant (Corollary 1).
4. When $n = d$, and $\Psi = 0.3d^{1/2}$, then the power should increase as $\sqrt{d}$ (Corollary 2).

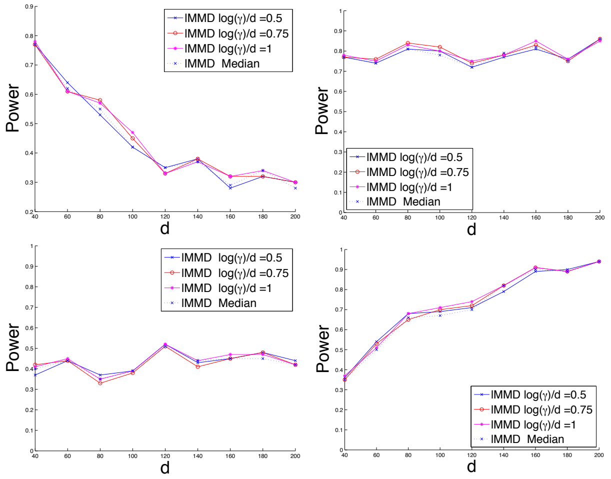

Figure 4: All plots show power vs d for different $\gamma \in \{\mathrm{median}, d^{0.5}, d^{0.75}, d\}$ for $d = 40$ to 200 in steps of 20. From top left to bottom right are the settings 1-4, with P, Q being Gaussians. The power is estimated over 100 repetitions at each d.

From Figure 4, we infer that the precise form of Theorem 1 (and Corollaries 1,2) is extremely accurate, even at small n and significantly larger d, including that it is independent of the bandwidth γ as predicted, as long as $\gamma = \Omega(\sqrt{d})$.

6 Conclusion

This paper has two main novelties - the first is to precisely characterize how a nonparametric two sample test, which is consistent in fixed dimensions against general alternatives, performs against a mean-shift alternative; the second is to perform the analysis in the significantly more difficult high-dimensional regime.

Future work involves understanding MMD_u, but the limiting distributions of general U-statistics are be difficult to ascertain in high dimensions. Another direction involves the study of *sparse* alternatives, where δ is sparse, as done by Cai et al. (2014). Lastly, minimax lower bounds are required to understand the tradeoffs involved between Ψ, d, n.

Acknowledgements

This research is supported in part by NSF grant IIS-1247658.

References

Anderson, Theodore W. *An introduction to multivariate statistical analysis*. Wiley, 1958.

Bai, Zhidong D and Saranadasa, Hewa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329, 1996.

Cai, Tony, Liu, Weidong, and Xia, Yin. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

Chen, Louis HY, Goldstein, Larry, and Shao, Qi-Man. *Normal approximation by Steins method*. Springer, 2010.

Chen, Song Xi and Qin, Ying-Li. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2): 808–835, apr 2010. doi: 10.1214/09-aos716. URL `http://dx.doi.org/10.1214/09-aos716`.

Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Hotelling, Harold. The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378, aug 1931. doi: 10.1214/aoms/1177732979. URL `http://dx.doi.org/10.1214/aoms/1177732979`.

Kariya, Takeaki. A robustness property of hotelling's t2-test. *The Annals of Statistics*, pp. 211–214, 1981.

Lehmann, Erich L and Romano, Joseph P. *Testing statistical hypotheses*. springer, 2006.

Lopes, M.E., Jacob, L., and Wainwright, M.J. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems 24*. MIT Press, 2011.

Lyons, R. Distance covariance in metric spaces. *Annals of Probability*, 41(5):3284–3305, 2013.

Ramdas, Aaditya, Reddi, Sashank J., Póczos, Barnabás, Singh, Aarti, and Wasserman, Larry A. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2015.

Rosenbaum, Paul R. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

Salaevskii, O.V. Minimax character of hotellings t2 test. i. In *Investigations in Classical Problems of Probability Theory and Mathematical Statistics*, pp. 74–101. Springer, 1971.

Schölkopf, Bernhard and Smola, A. J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Serfling, Robert J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Simaika, JB. On an optimum property of two important statistical tests. *Biometrika*, pp. 70–80, 1941.

Srivastava, Muni S. and Du, Meng. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402, mar 2008. doi: 10.1016/j.jmva.2006.11.002. URL `http://dx.doi.org/10.1016/j.jmva.2006.11.002`.

Székely, Gábor J and Rizzo, Maria L. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

Székely, G.J. and Rizzo, M.L. The distance correlation t-test of independence in high dimension. *J. Multivariate Analysis*, 117:193–213, 2013.