A Proofs of Propositions 1,2,3.

Before we look at the MMD calculations in various cases, we prove the following useful characterization of MMD for translation invariant kernels like the Gaussian and Laplace kernels.

Lemma 1. *For translation invariant kernels, there exists a pdf s such that*

$$\mathrm{MMD}^2(p,q) = \int s(w)|\Phi_p(w) - \Phi_q(w)|^2 dw,$$

where Φ_p, Φ_q denote the characteristic functions of p, q respectively.

Proof. From definition of MMD^2, we have

$$\mathrm{MMD}^2(p,q) = \int\limits_{x,x'} k(x,x')p(x)p(x')dxdx' + \int\limits_{x,x'} k(x,x')q(x)q(x')dxdx' - 2\int\limits_{x,x'} k(x,x')p(x)q(x')dxdx'.$$

From Bochner's theorem (see (Rudin 1962)) for translation invariant kernels, we know $k(x,x') = \int_w s(w)e^{iw^\top x}e^{-iw^\top x'}dw$ where s is the fourier transform of the kernel. Substituting the above equality in the definition of MMD^2, we have the required result. □

Proof of Proposition 1

Proof. Since Gaussian kernel is a translation invariant kernel, we can use Lemma 1 to derive the MMD^2 in this case. It is well-known that the Fourier transform $s(w)$ of Gaussian kernel is Gaussian distribution. Substituting the characteristic function of normal distribution in Lemma 1, we have

$$\begin{aligned}
\mathrm{MMD}^2(p,q) &= \int_w \left(\gamma^2/2\pi\right)^{d/2} \exp\left(-\gamma^2\|w\|^2/2\right) \left|\exp(i\mu_1^\top w - w^\top\Sigma w/2) - \exp(i\mu_1^\top w - w^\top\Sigma w/2)\right|^2 dw \\
&= \left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top\Sigma w\right) \exp\left(-\gamma^2\|w\|^2/2\right) \left|\exp(i\mu_1^\top w) - \exp(i\mu_2^\top w)\right|^2 dw \\
&= \left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top(\Sigma + \gamma^2 I/2)w\right) \left(2 - \exp\left(-i(\mu_1-\mu_2)^\top w\right) - \exp\left(-i(\mu_2-\mu_1)^\top w\right)\right) dw \\
&= 2\left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top(\Sigma + \gamma^2 I/2)w\right) \left(1 - \exp\left(-i(\mu_1-\mu_2)^\top w\right)\right) dw \qquad (2)
\end{aligned}$$

The third step follows from definition of complex conjugate. In what follows, we do the following change of variable $u = (\Sigma + \gamma^2 I/2)^{1/2}w$. Consider the following term:

$$\begin{aligned}
&\int_w \exp\left(-w^\top(\Sigma + \gamma^2 I/2)w\right) \exp\left(-i(\mu_1-\mu_2)^\top w\right) dw \\
&\quad = \int_u \exp -\left(u^\top u + i(\mu_1-\mu_2)^\top(\Sigma + \gamma^2 I/2)^{-1/2}u\right) |\Sigma + \gamma^2 I/2|^{-1/2} du \\
&\quad = |\Sigma + \gamma^2 I/2|^{-1/2} \exp(-(\mu_1-\mu_2)^\top(\Sigma + \gamma^2 I/2)^{-1}(\mu_1-\mu_2)/4) \times \\
&\qquad\qquad \int_u \exp -\left(\|u - i(\Sigma + \gamma^2 I/2)^{-1/2}(\mu_1-\mu_2)/2\|^2\right) du \\
&\quad = \pi^{d/2}|\Sigma + \gamma^2 I/2|^{-1/2} \exp(-(\mu_1-\mu_2)^\top(\Sigma + \gamma^2 I/2)^{-1}(\mu_1-\mu_2)/4)
\end{aligned}$$

The second step follows from well-known theory of change of variables (see Theorem 263D of (Fremlin 2000)). By substituting the above equality in Equation 2, we get the required result. □

Proof of Proposition 2

Before we delve into the details of the result, we prove the following useful propositions.

Proposition 4. *Let $\sigma, \gamma \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}$. Suppose $\gamma \neq \sigma$, then we have,*

$$\int\limits_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\gamma}\right) \exp\left(-\frac{|x|}{\sigma}\right) dx \quad = \quad \frac{e^{-|\lambda|/\sigma}}{1/\gamma + 1/\sigma} + \frac{e^{-|\lambda|/\gamma}}{1/\sigma - 1/\gamma} - \frac{e^{-|\lambda|/\sigma}}{1/\sigma - 1/\gamma} + \frac{e^{-|\lambda|/\gamma}}{1/\gamma + 1/\sigma}$$

and when $\gamma = \sigma$, we have,

$$\int\limits_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\sigma}\right) \exp\left(-\frac{|x|}{\sigma}\right) dx \quad = \quad \frac{e^{-|\lambda|/\sigma}}{1/\gamma + 1/\sigma} + |\lambda|e^{-|\lambda|/\sigma} + \frac{e^{-|\lambda|/\gamma}}{1/\gamma + 1/\sigma}$$

Proof. We show this when $\lambda \leq 0$ as an example proof:

$$\begin{aligned}
\int_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\gamma}\right)\exp\left(-\frac{|x|}{\sigma}\right)dx &= \int_{-\infty}^{\lambda}\exp\left(\frac{x-\lambda}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx + \int_{\lambda}^{0}\exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx \\
&\quad + \int_{0}^{\infty}\exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(-\frac{x}{\sigma}\right)dx \\
&= \frac{e^{-\lambda/\gamma}e^{\lambda/\sigma+\lambda/\gamma}}{1/\gamma+1/\sigma} + \frac{e^{-\lambda/\gamma}(1-e^{-\lambda/\gamma+\lambda/\sigma})}{1/\sigma-1/\gamma} + \frac{e^{\lambda/\gamma}}{1/\gamma+1/\sigma}
\end{aligned}$$

Also, when $\gamma = \sigma$, we obtain the same expression for the first and last terms. However, the middle term has the following constant integrand, thereby, leading to the required expression.

$$\int_{\lambda}^{0}\exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx = |\lambda|e^{-|\lambda|/\sigma}.$$

□

Proposition 5. *Let $\sigma, \gamma \in \mathbb{R}^+$ and $\mu \in \mathbb{R}$. Then we have,*

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{|x-x'|}{\gamma}\right)\frac{1}{4\sigma^2}\exp\left(-\frac{|x-\mu|}{\sigma}\right)\exp\left(-\frac{|x'|}{\sigma}\right)dxdx' \\
&= -\frac{1}{2}e^{-|\mu|/\sigma}\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right) + \frac{1}{1-\psi^2}\left(-\frac{\psi e^{-|\mu|/\sigma}}{1-\psi^2}+\frac{e^{-|\mu|/\gamma}}{1-\psi^2}\right) \\
&= -\frac{\mu^2}{4\sigma\gamma(1+\psi)^2}+\frac{2+\psi}{2(1+\psi)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right)
\end{aligned}$$

where $\psi = \sigma/\gamma$.

Proof. We first integrate with respect to x' using the Proposition 4 to get

$$\frac{1}{4\sigma^2}\int_{-\infty}^{\infty}\left(\frac{e^{-|x|/\sigma}}{1/\gamma+1/\sigma}+\frac{e^{-|x|/\gamma}}{1/\sigma-1/\gamma}-\frac{e^{-|x|/\sigma}}{1/\sigma-1/\gamma}+\frac{e^{-|x|/\gamma}}{1/\gamma+1/\sigma}\right)\exp\left(-\frac{|x-\mu|}{\sigma}\right)dx$$

We then integrate these terms once again using both parts of Proposition 4 to get the first equality. We simplify the second equation in the following manner:

$$\begin{aligned}
&-\frac{1}{2}e^{-|\mu|/\sigma}\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right)+\frac{1}{1-\psi^2}\left(-\frac{\psi e^{-|\mu|/\sigma}}{1-\psi^2}+\frac{e^{-|\mu|/\gamma}}{1-\psi^2}\right) \\
=\ & -\frac{1}{2}\left(1-\frac{|\mu|}{\sigma}+\frac{|\mu|^2}{2\sigma^2}\right)\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right)+\frac{1}{1-\psi^2}\left(-\frac{(\sigma/\gamma-|\mu|/\gamma+\mu^2/2\sigma\gamma)}{1-\psi^2}+\frac{1-|\mu|/\gamma+\mu^2/2\gamma^2}{1-\psi^2}\right) \\
&\quad +O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
=\ & -\frac{1}{2(1-\psi^2)}\left(\psi-\frac{\mu^2}{2\sigma\gamma}+\frac{|\mu|^3}{2\sigma^2\gamma}\right)+\frac{1}{(1-\psi^2)^2}\left(1-\psi-\frac{\mu^2}{2\sigma\gamma}+\frac{\mu^2}{2\gamma^2}\right) \\
&\quad +O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
=\ & -\frac{1}{2(1-\psi^2)}\left(\psi-\frac{\mu^2}{2\sigma\gamma}\right)+\frac{(1-\mu^2/2\sigma\gamma)(1-\psi)}{(1-\psi^2)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
=\ & \frac{1}{1-\psi^2}\left(-\frac{\psi}{2}+\frac{1}{2}\frac{\mu^2}{2\sigma\gamma}\right)+\frac{1}{1-\psi^2}\left(\frac{1}{1+\psi}-\frac{\mu^2}{(1+\psi)2\sigma\gamma}\right)+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
=\ & -\frac{\mu^2}{4\sigma\gamma(1+\psi)^2}+\frac{2+\psi}{2(1+\psi)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right)
\end{aligned}$$

□

Proof (Proposition 2). Recall that we use Laplace kernel, i.e., $k(x, x') = \exp(-\|x - x'\|_1/\gamma)$. By using the definition of MMD^2, we have

$$\mathrm{MMD}^2 = \int_{x,x'} (p(x)p(x') + q(x)q(x') - 2p(x)q(x'))k(x, x')dxdx'. \tag{3}$$

Consider the term $\int_{x,x'} p(x)q(x')k(x, x')dxdx'$. The other terms can be calculated in a similar manner. Let $\psi = \sigma/\gamma$ and $\beta = (1 + \psi/2)/(1 + \psi)^2$. We have,

$$\begin{aligned}\int_{x,x'} p(x)q(x')k(x, x')dxdx' &= \prod_{i=1}^{d} \int_{x_i,x'_i} \exp\left(-\frac{|x - x'|}{\gamma}\right) \frac{1}{4\sigma^2} \exp\left(-\frac{|x - \mu|}{\sigma}\right) \exp\left(-\frac{|x'|}{\sigma}\right) dx_i dx)i' \\ &= \prod_{i=1}^{d} \beta \left(1 - \frac{\mu_i^2}{4\beta\sigma\gamma(1 + \psi)^2} + O\left(\frac{|\mu_i|^3}{\beta\sigma^2\gamma(1 - \psi^2)^2}\right) - O\left(\frac{|\mu_i|^3}{\beta\gamma^3(1 - \psi^2)^2}\right)\right) \\ &= \beta^d \left(1 - \frac{\|\mu\|^2}{4\beta\sigma\gamma(1 + \psi)} + O\left(\frac{|\mu_i|^3}{\beta\sigma^2\gamma(1 - \psi^2)^2}\right) - O\left(\frac{|\mu_i|^3}{\beta\gamma^3(1 - \psi^2)^2}\right)\right)\end{aligned}$$

The first step follows from the fact that both Laplace kernel and Laplace distribution decompose over the coordinates. The second step follows from Proposition 5. Substituting the above expression in Equation 3, we get,

$$\mathrm{MMD}^2 = \frac{\beta^{d-1}\|\mu\|^2}{2\sigma\gamma(1 + \psi)} - O\left(\frac{\beta^{d-1}\|\mu\|_3^3}{\sigma^2\gamma(1 - \psi^2)^2}\right) + O\left(\frac{\beta^{d-1}\|\mu\|_3^3}{\gamma^3(1 - \psi^2)^2}\right).$$

□

Proof of Proposition 3

Suppose $P = \otimes_{i=1}^d N(0, \sigma^2) \otimes N(0, a^2)$ and $Q = \otimes_{i=1}^d N(0, \sigma^2) \otimes N(0, b^2)$. If a, b are of the same order as σ then the median heuristic will still pick $\gamma \approx \sigma\sqrt{d}$ for bandwidth γ of the Gaussian kernel. First we note that for distributions with the same mean, by Taylor's theorem,

$$\begin{aligned}KL(P, Q) &= \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0 - d - \log(\det \Sigma_0)/\det \Sigma_1)) = \frac{1}{2}(a^2/b^2 - 1 - \log(a^2/b^2)) \\ &\approx \frac{(a^2/b^2 - 1)^2}{4}\end{aligned}$$

The MMD^2 can be derived (approximated using $(1 + x)^n \approx 1 + nx$ for small x) as

$$\begin{aligned}& \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{\sqrt{1 + 4a^2/\gamma^2}} + \frac{1}{\sqrt{1 + 4b^2/\gamma^2}} - \frac{2}{\sqrt{1 + 2(a^2 + b^2)/\gamma^2}}\right) \\ \approx\ & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{1 + 2a^2/\gamma^2} + \frac{1}{1 + 2b^2/\gamma^2} - \frac{2}{1 + (a^2 + b^2)/\gamma^2}\right) \\ \approx\ & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{\sqrt{1 + 2a^2/\gamma^2}} - \frac{1}{\sqrt{1 + 2b^2/\gamma^2}}\right)^2 \\ \approx\ & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left((1 - a^2/\gamma^2) - (1 - b^2/\gamma^2)\right)^2 \\ =\ & \frac{b^4/\gamma^4}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} (a^2/b^2 - 1)^2\end{aligned}$$

If γ is chosen by the median heuristic (optimal in this case), we see that this is smaller than KL by $\sigma^4 d^2 e/b^4$. If it is chosen as constant, it can be exponentially smaller than KL.

B Verifying accuracy of approximate MMDs calculated in Propositions 1,2,3.

In the proofs and corollaries of derivations of MMD in Propositions 1,2,3, we used many Taylor approximations in order to get a more interpretable formula. Here we show that our approximate formulae, while being interpretable, are also very accurate.

We provide empirical results demonstrating the quality of the approximations used in Section 4. In particular, we compare the estimated value of the MMD using *large* sample size (so that the sample MMD is a very good estimate of population MMD) and the approximations provided in Section 4. As observed in Figure 8, the approximations are quite close to the estimated value, thereby validating the quality of our approximations.

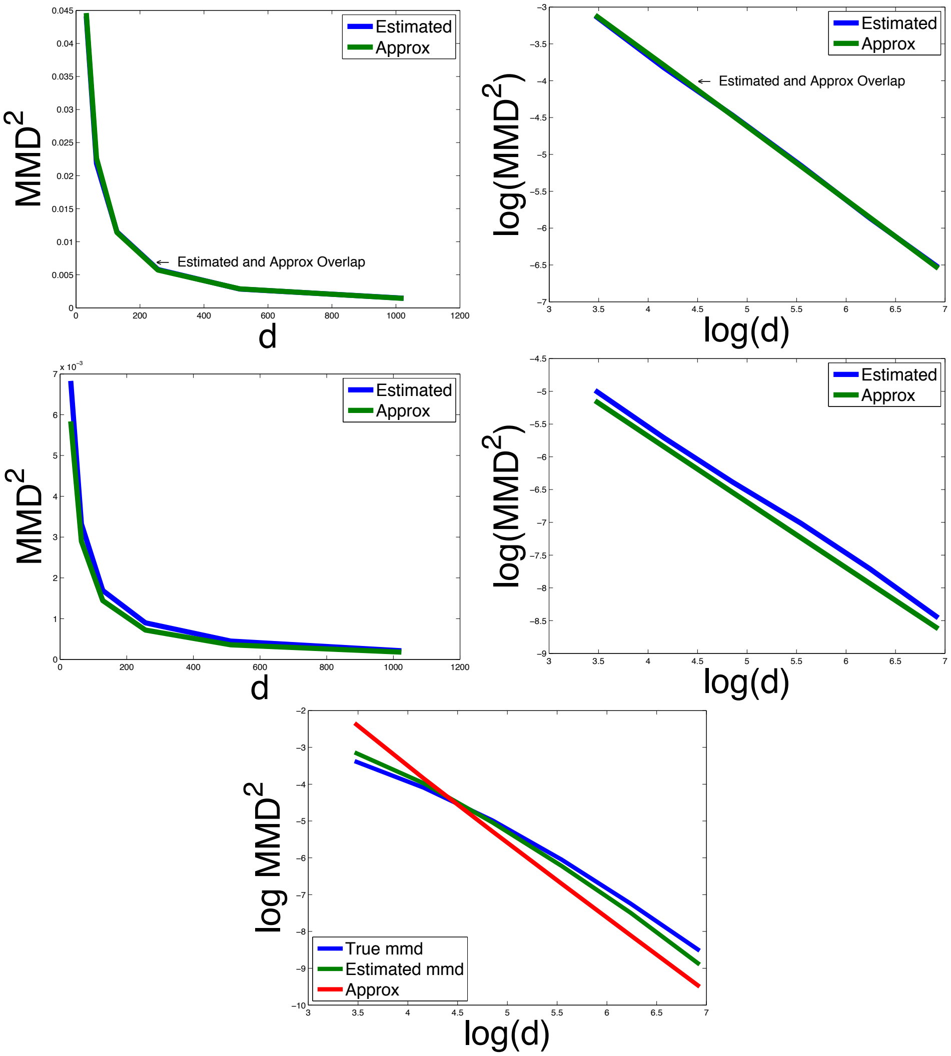

Figure 8: Top left: MMD vs d, for Gaussian distributions and Gaussian kernel with optimal $\sigma\sqrt{d}$ bandwidth, as estimated from data and approximated by formula. Top right: same but for Log(MMD). Middle left: MMD vs d, for Laplace kernel with optimal σd bandwidth, estimated from data and approximated by formula. Middle right: same but for Log(MMD). The Log Plots also show the right scaling that decays as $1/d$ with the right choice of bandwidth. Bottom: Log(MMD) vs d, for Gaussian kernel with optimal $\sigma\sqrt{d}$ bandwidth, for Gaussians with same mean and different variances. The straight line is our final approximation in the theorem. The other two are the true MMD by formula, and the MMD from data.

C Biased MMD for Gaussian Distribution

In the previous sections, we provided results for unbiased MMD estimator and empirically proved that the power of the test based on the estimator decreases with increasing dimension. We report results for the biased MMD estimator in this section and show that it exhibits similar behavior.

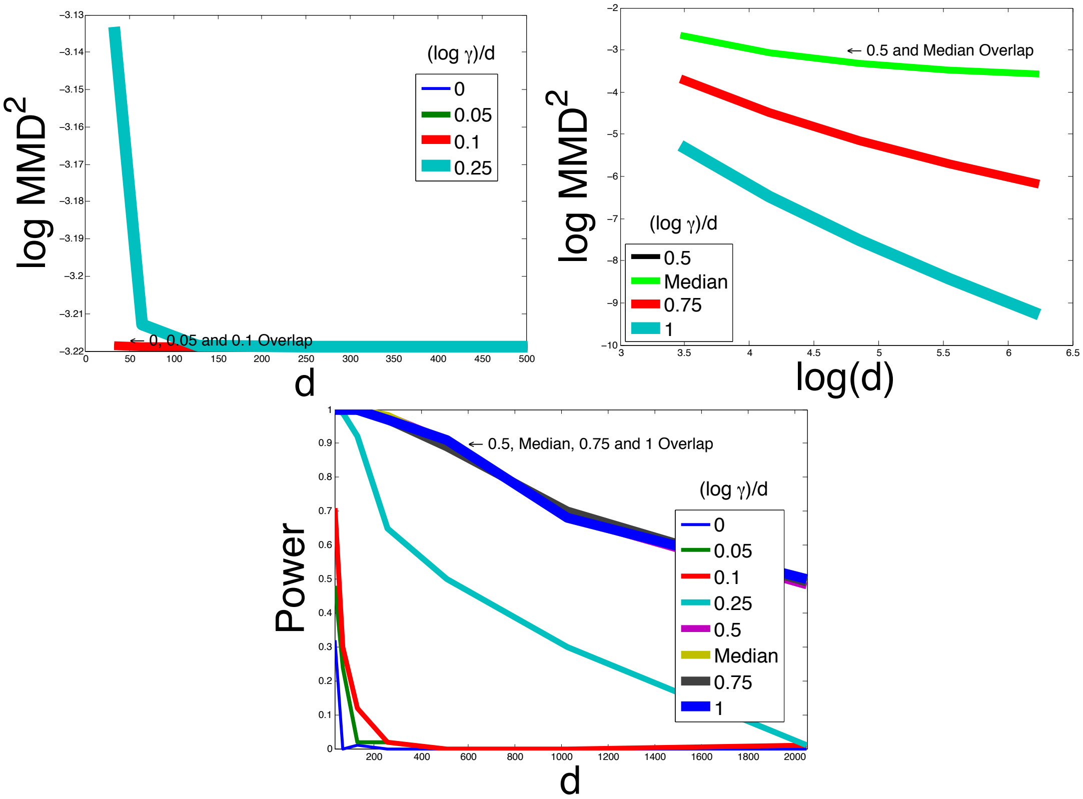

Figure 9: Plots for Biased MMD with Gaussian kernel, when the data is drawn from two Gaussians with $\sigma^2 = 1$ and constant mean separation $\|\mu_1 - \mu_2\|^2 = 1$. With respect to the selection of bandwidth γ, the power of Biased MMD has similar behavior as Unbiased MMD.

As seen in Figure 9, the power of the biased MMD decreases in exactly the same fashion as unbiased MMD. We also observed similar behavior with other examples.