A. Theoretical analysis

Appendix A.2 proves the main results under some assumptions about the kernel parameterization, using intermediate results about uniform convergence of our estimators in Appendix A.3. Appendix A.4 then shows that these assumptions hold for different settings of kernel learning.

A.1. Preliminaries

Given a kernel k_ω and sample sets $\{X_i\}_{i=1}^n \sim \mathbb{P}^n$, $\{Y_i\}_{i=1}^n \sim \mathbb{Q}^n$, define the $n \times n$ matrix

$$H_{ij}^{(\omega)} = k_\omega(X_i, X_j) + k_\omega(Y_i, Y_j) - k_\omega(X_i, Y_j) - k_\omega(X_j, Y_i);$$

we will often omit ω when it is clear from context. The U-statistic estimator of the squared MMD (2) is

$$\hat{\eta}_\omega = \frac{1}{n(n-1)} \sum_{i \neq j} H_{ij}.$$

The squared MMD is $\eta_\omega = \mathbb{E}[H_{12}]$. The variance of $\hat{\eta}_\omega$ is given by Lemma 10.

Lemma 10. *For a fixed kernel k_ω and random sample sets $\{X_i\}_{i=1}^n$, $\{Y_i\}_{i=1}^n$, we have*

$$\mathrm{Var}[\hat{\eta}_\omega] = \frac{4(n-2)}{n(n-1)} \xi_1^{(\omega)} + \frac{2}{n(n-1)} \xi_2^{(\omega)} = \frac{4}{n} \xi_1^{(\omega)} + \frac{2\xi_2^{(\omega)} - 4\xi_1^{(\omega)}}{n(n-1)}, \tag{8}$$

where

$$\xi_1^{(\omega)} = \mathbb{E}\left[H_{12}^{(\omega)} H_{13}^{(\omega)}\right] - \mathbb{E}\left[H_{12}^{(\omega)}\right]^2, \qquad \xi_2^{(\omega)} = \mathbb{E}\left[\left(H_{12}^{(\omega)}\right)^2\right] - \mathbb{E}\left[H_{12}^{(\omega)}\right]^2.$$

Thus as $n \to \infty$,

$$n \,\mathrm{Var}[\hat{\eta}_\omega] \to 4\xi_1^{(\omega)} =: \sigma_\omega^2.$$

Proof. Let U denote the pair (X, Y), and $h_\omega(U, U') = k_\omega(X, X') + k_\omega(Y, Y') - k_\omega(X, Y') - k_\omega(X', Y)$, so that $H_{ij}^{(\omega)} = h_\omega(U_i, U_j)$. Via Lemma A in Section 5.2.1 of Serfling (1980), we know that (8) holds with

$$\begin{aligned}
\xi_1^{(\omega)} &= \mathrm{Var}_U\left[\mathbb{E}_{U'}\left[h_\omega(U, U')\right]\right] \\
&= \mathbb{E}_U\left[\mathbb{E}_{U'}\left[h_\omega(U, U')\right] \mathbb{E}_{U''}\left[h_\omega(U, U'')\right]\right] - \mathbb{E}_U\left[\mathbb{E}_{U'}[h_\omega(U, U')]\right]^2 \\
&= \mathbb{E}[H_{12}^{(\omega)} H_{13}^{(\omega)}] - \mathbb{E}[H_{12}^{(\omega)}]^2
\end{aligned}$$

and

$$\xi_2 = \mathrm{Var}_{U,U'}\left[h_\omega(U, U')\right] = \mathbb{E}\left[\left(H_{12}^{(\omega)}\right)^2\right] - \mathbb{E}\left[H_{12}^{(\omega)}\right]^2. \qquad \square$$

We use a V-statistic estimator (5) for σ_ω^2:

$$\hat{\sigma}_\omega^2 = 4\left(\frac{1}{n}\sum_{i=1}^n \left(\frac{1}{n}\sum_{j=1}^n H_{ij}^{(\omega)}\right)^2 - \left(\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n H_{ij}^{(\omega)}\right)^2\right).$$

As a V-statistic, $\hat{\sigma}_\omega^2$ is biased. In fact, Sutherland et al. (2017) and Sutherland (2019) provide an unbiased estimator of $\mathrm{Var}[\hat{\eta}_\omega]$ – including the terms of order $\frac{1}{n(n-1)}$. Although this estimator takes the same quadratic time to compute as (5), it contains many more terms, which are cumbersome both for implementation and for analysis. (5) is also marginally more convenient in that it is always at least nonnegative. As we show in Lemma 18, the amount of bias is negligible as n increases. In practice, we expect the difference to be unimportant – or the V-statistic may in fact be beneficial, since underestimating σ^2 harms the estimate of η/σ^2 more than overestimating it does.

Similarly, although we use the U-statistic estimator (2), it would be very similar to use the biased estimator $n^{-2}\sum_{ij} H_{ij}$, or the minimum variance unbiased estimator $n^{-1}(n-1)^{-1}\sum_{i\neq j}(k(X_i, X_j) + k(Y_i, Y_j)) - 2n^{-2}\sum_{ij} k(X_i, Y_J)$. Showing comparable concentration behavior to Proposition 15 is trivially different, and in fact it is also not difficult to show σ_ω^2 is the same for all three estimators (up to lower-order terms).

A.2. Main results

We will require the following assumptions. These are fairly agnostic as to the kernel form; Appendix A.4.2 shows that these assumptions hold (and gives the constants) for the kernels (1) we use in the paper.

(A) The kernels k_ω are uniformly bounded:

$$\sup_{\omega\in\Omega}\sup_{x\in\mathcal{X}} k_\omega(x,x) \le \nu.$$

For the kernels we use in practice, $\nu = 1$.

(B) The possible kernel parameters ω lie in a Banach space of dimension D. Furthermore, the set of possible kernel parameters Ω is bounded by R_ω, $\Omega \subseteq \{\omega \mid \|\omega\| \le R_\Omega\}$.

Appendix A.4.2 builds this space and its norm for the kernels we use in the paper.

(C) The kernel parameterization is Lipschitz: for all $x, y \in \mathcal{X}$ and $\omega, \omega' \in \Omega$,

$$|k_\omega(x,y) - k_{\omega'}(x,y)| \le L_k\|\omega - \omega'\|.$$

Proposition 23 in Appendix A.4.2 gives an expression for L_k for the kernels we use in the paper.

We will first show the main results under these general assumptions, using uniform convergence results shown in Appendix A.3, then show Assumptions **(B)** and **(C)** for particular kernels in Appendix A.4.2.

Theorem 11. *Under Assumptions (A) to (C), let $\bar\Omega_s \subseteq \Omega$ be the set of kernel parameters for which $\sigma_\omega^2 \ge s^2$, and assume $\nu \ge 1$. Take $\lambda = n^{-1/3}$. Then, with probability at least $1-\delta$,*

$$\sup_{\omega\in\bar\Omega_s}\left|\frac{\hat\eta_\omega}{\hat\sigma_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}\right| \le \frac{2\nu}{s^2 n^{1/3}}\left(\frac{1}{s} + \frac{2304\nu^2}{\sqrt{n}} + \left[\frac{4s}{n^{1/6}} + 1024\nu\right]\left[L_k + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}\right]\right),$$

and thus, treating ν as a constant,

$$\sup_{\omega\in\bar\Omega_s}|\frac{\hat\eta_\omega}{\hat\sigma_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}| = \tilde{\mathcal{O}}_P\left(\frac{1}{s^2 n^{1/3}}\left[\frac{1}{s} + L_k + \sqrt{D}\right]\right).$$

Proof. Let $\sigma_{\omega,\lambda}^2 := \sigma_\omega^2 + \lambda$. Using $|\hat\eta_\omega| \le 4\nu$, we begin by decomposing

$$\begin{aligned}
\sup_{\omega\in\bar\Omega_s}|\frac{\hat\eta_\omega}{\hat\sigma_{\omega,\lambda}} - \frac{\eta_\omega}{\sigma_\omega}| &\le \sup_{\omega\in\bar\Omega_s}|\frac{\hat\eta_\omega}{\hat\sigma_{\omega,\lambda}} - \frac{\hat\eta_\omega}{\sigma_{\omega,\lambda}}| + \sup_{\omega\in\bar\Omega_s}|\frac{\hat\eta_\omega}{\sigma_{\omega,\lambda}} - \frac{\hat\eta_\omega}{\sigma_\omega}| + \sup_{\omega\in\bar\Omega_s}|\frac{\hat\eta_\omega}{\sigma_\omega} - \frac{\eta_\omega}{\sigma_\omega}| \\
&= \sup_{\omega\in\bar\Omega_s}|\hat\eta_\omega|\frac{1}{\hat\sigma_{\omega,\lambda}}\frac{1}{\sigma_{\omega,\lambda}}\frac{|\hat\sigma_{\omega,\lambda}^2 - \sigma_{\omega,\lambda}^2|}{\hat\sigma_{\omega,\lambda} + \sigma_{\omega,\lambda}} + \sup_{\omega\in\bar\Omega_s}|\hat\eta_\omega|\frac{1}{\sigma_{\omega,\lambda}}\frac{1}{\sigma_\omega}\frac{|\sigma_{\omega,\lambda}^2 - \sigma_\omega^2|}{\sigma_{\omega,\lambda} + \sigma_\omega} + \sup_{\omega\in\bar\Omega_s}\frac{1}{\sigma_\omega}|\hat\eta_\omega - \eta_\omega| \\
&\le \sup_{\omega\in\bar\Omega_s}\frac{4\nu}{\sqrt{\lambda}\,s\,(s+\sqrt{\lambda})}|\hat\sigma_\omega^2 - \sigma_\omega^2| + \frac{4\nu\lambda}{\sqrt{s^2+\lambda}\,s\,\left(\sqrt{s^2+\lambda}+s\right)} + \sup_{\omega\in\bar\Omega_s}\frac{1}{s}|\hat\eta_\omega - \eta_\omega| \\
&\le \frac{4\nu}{s^2\sqrt{\lambda}}\sup_{\omega\in\Omega}|\hat\sigma_\omega^2 - \sigma_\omega^2| + \frac{2\nu}{s^3}\lambda + \frac{1}{s}\sup_{\omega\in\Omega}|\hat\eta_\omega - \eta_\omega|.
\end{aligned}$$

Propositions 15 and 16 show uniform convergence of $\hat\eta_\omega$ and $\hat\sigma_\omega^2$, respectively. Thus, with probability at least $1-\delta$, the error is at most

$$\frac{2\nu}{s^3}\lambda + \left[\frac{8\nu}{s\sqrt{n}} + \frac{1792\nu}{\sqrt{n}s^2\sqrt{\lambda}}\right]\sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)} + \left[\frac{8}{s\sqrt{n}} + \frac{2048\nu^2}{\sqrt{n}s^2\sqrt{\lambda}}\right]L_k + \frac{4608\nu^3}{s^2 n\sqrt{\lambda}}.$$

Taking $\lambda = n^{-1/3}$ gives

$$\frac{2\nu}{s^3 n^{1/3}} + \left[\frac{8\nu}{s\sqrt{n}} + \frac{1792\nu}{s^2 n^{1/3}}\right]\sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)} + \left[\frac{8}{s\sqrt{n}} + \frac{2048\nu^2}{s^2 n^{1/3}}\right]L_k + \frac{4608\nu^3}{s^2 n^{5/6}}.$$

Using $1 \leq \nu$, $1792 < 2048$, we can get the slightly simpler upper bound

$$\frac{2\nu}{s^3 n^{1/3}} + \left[\frac{8\nu}{s\sqrt{n}} + \frac{2048\nu^2}{s^2 n^{1/3}}\right]\left[L_k + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}\right] + \frac{4608\nu^3}{s^2 n^{5/6}}. \qquad \square$$

It is worth noting that, if we are particularly concerned about the s dependence, we can make some slightly different choices in the decomposition to improve the dependence on s while worsening the rate with n.

Corollary 12. *In the setup of Theorem 11, additionally assume that there is a unique population maximizer ω^* of J from (3), i.e. for each $t > 0$ we have*

$$\sup_{\omega \in \bar{\Omega}_s : \|\omega - \omega^*\| \geq t} J(\mathbb{P}, \mathbb{Q}; k_\omega) < J(\mathbb{P}, \mathbb{Q}; k_{\omega^*}).$$

For each n, let $S_{\mathbb{P}}^{(n)}$ and $S_{\mathbb{Q}}^{(n)}$ be sequences of sample sets of size n, let $\hat{J}_n(\omega)$ denote $J_{\lambda=n^{-1/3}}(S_{\mathbb{P}}^{(n)}, S_{\mathbb{Q}}^{(n)}; k_\omega)$, and take $\hat{\omega}_n^*$ to be a maximizer of $\hat{J}_n(\omega)$.[7] Then $\hat{\omega}_n^*$ converges in probability to ω^*.

Proof. By Theorem 11, $\sup_{\omega \in \bar{\Omega}_s} |\hat{J}_n(\omega) - J(\omega)| \xrightarrow{P} 0$. Then the result follows by Theorem 5.7 of Van der Vaart (2000). $\square$

Corollary 13. *In the setup of Theorem 11, suppose we use n sample points to select a kernel $\hat{\omega}_n \in \arg\max_{\omega \in \bar{\Omega}_s} \hat{J}_\lambda(\omega)$ and m sample points to run a test of level α. Let $r_{\hat{\omega}_n}^{(m)}$ denote the rejection threshold for a test with that kernel of size m. Define $J^* := \sup_{\omega \in \bar{\Omega}_s} J(\omega)$, and constants C, C', C'', N_0 depending on ν, L_k, D, R_Ω and s. For any $n \geq N_0$, with probability at least $1 - \delta$, this test procedure has power*

$$\Pr\left(m\,\hat{\eta}_{\hat{\omega}_n} > r_{\hat{\omega}_n}^{(m)}\right) \geq \Phi\left(\sqrt{m}J^* - C\frac{\sqrt{m}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}} - C'\sqrt{\log\frac{1}{\alpha}}\right) - \frac{C''}{\sqrt{m}}.$$

Proof. Let $\hat{\omega}_n \in \arg\max_{\omega \in \bar{\Omega}_s} \hat{J}_\lambda(\omega)$. By Theorem 11, there are some N_0, C depending on ν, L_k, D, R_Ω, and s such that as long as $n \geq N_0$, with probability at least $1 - \delta$ it holds that

$$\sup_{\omega \in \bar{\Omega}_s} |J_\lambda(\omega) - J(\omega)| \leq \tfrac{1}{2} C n^{-\frac{1}{3}} \sqrt{\log\frac{n}{\delta}} =: \epsilon_n.$$

Assume for the remainder of this proof that this event holds. Letting $\omega^* \in \arg\max J(\omega)$, we know because $\hat{\omega}_n$ maximizes $\hat{J}_\lambda$ that $\hat{J}_\lambda(\hat{\omega}_n) \geq \hat{J}_\lambda(\omega^*)$. Using uniform convergence twice,

$$J(\hat{\omega}_n) \geq \hat{J}_\lambda(\hat{\omega}_n) - \epsilon_n \geq \hat{J}_\lambda(\omega^*) - \epsilon_n \geq (J(\omega^*) - \epsilon_n) - \epsilon_n = J^* - 2\epsilon_n.$$

Now, although Proposition 2 establishes that $r_\omega^{(m)} \to r_\omega$ and it is even known (Korolyuk & Borovskikh, 1988, Theorem 5) that $|r_\omega^{(m)} - r_\omega|$ is $o(1/\sqrt{m})$, the constant in that convergence will depend on the choice of ω in an unknown way. It's thus simpler to use the very loose but uniform (McDiarmid-based) bound given by Corollary 11 of Gretton et al. (2012a), which implies $r_\omega^{(m)} \leq 4\nu\sqrt{\log(\alpha^{-1})m}$ no matter the choice of ω.

We will now need a more precise characterization of the power than that provided by the central limit theorem of Proposition 2. Callaert & Janssen (1978) provide such a result, a Berry-Esseen bound on U-statistic convergence: there is some absolute constant $C'_{BS} = 2^3 4^3 C_{BS}$ such that

$$\sup_t \left|\Pr_{\mathfrak{H}_1}\left(\sqrt{m}\frac{\hat{\eta}_\omega - \eta_\omega}{\sigma_\omega^2} \leq t\right) - \Phi(t)\right| \leq \frac{C'_{BS}\,\mathbb{E}|H_{12}|^3}{(\sigma_\omega/2)^3\sqrt{m}} \leq \frac{C_{BS}\nu^3}{\sigma_\omega^3\sqrt{m}}.$$

[7] In fact, it suffices for the $\hat{\omega}_n^*$ to only approximately maximize $\hat{J}_n$, as long as their suboptimality is $o_P(1)$.

Letting $r_\omega^{(m)}$ be the appropriate rejection threshold for k_ω with m samples, the power of a test with kernel k_ω is

$$\begin{aligned}\Pr\left(m\hat{\eta}_\omega > r_\omega^{(m)}\right) &= \Pr\left(\sqrt{m}\frac{\hat{\eta}_\omega - \eta_\omega}{\sigma_\omega} > \frac{r_\omega^{(m)}}{\sqrt{m}\sigma_\omega} - \sqrt{m}\frac{\eta_\omega}{\sigma_\omega}\right)\\ &\geq \Phi\left(\sqrt{m}J(\omega) - \frac{r_\omega^{(m)}}{\sqrt{m}\sigma_\omega}\right) - \frac{C_{BS}\nu^3}{\sigma_\omega^3\sqrt{m}}\\ &\geq \Phi\left(\sqrt{m}J(\omega) - \frac{r_\omega^{(m)}}{s\sqrt{m}}\right) - \frac{C''}{\sqrt{m}},\end{aligned}$$

using a new constant $C'' := C_{BS}\nu^3/s^3$. Combining the previous results on $J(\hat{\omega}_n)$ and $r_{\hat{\omega}_n}^{(m)}$ yields the claim. □

Corollary 14. *In the setup of Corollary 13, suppose we are given N data points to divide between n training points and $m = N - n$ testing points, and $\delta < 0.22$ is fixed. Ignoring the Berry-Esseen convergence term outside of Φ, the asymptotic power upper bound*

$$\Phi\left(\sqrt{m}J^* - C\frac{\sqrt{m}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}} - C'\sqrt{\log\frac{1}{\alpha}}\right)$$

is maximized only when, as other quantities remain constant,

$$\lim_{N\to\infty}\frac{n}{\left(\frac{C}{\sqrt{3}J^*}N\sqrt{\log N}\right)^{\frac{3}{4}}} = 1.$$

Proof. Because the C' term is constant, we wish to choose

$$\arg\max_{0<n<N}\frac{J^*}{C}\sqrt{N-n} - \frac{\sqrt{N-n}}{n^{\frac{1}{3}}}\sqrt{\log\frac{n}{\delta}}.$$

Clearly neither endpoint is optimal. Relaxing n to be real-valued, the optimum must be achieved at a stationary point, where

$$\frac{-J^*}{2C\sqrt{N-n}} + \frac{\sqrt{\log\frac{n}{\delta}}}{2\sqrt{N-n}\,n^{\frac{1}{3}}} + \frac{1}{3}\sqrt{N-n}\,n^{-\frac{4}{3}}\sqrt{\log\frac{n}{\delta}} - \frac{1}{2}\sqrt{N-n}\,n^{-\frac{4}{3}}\left(\log\frac{n}{\delta}\right)^{-\frac{1}{2}} = 0.$$

Multiplying by $2\sqrt{N-n}\,n^{\frac{4}{3}}\sqrt{\log\frac{n}{\delta}}$ and rearranging, we get that a stationary point is achieved exactly when

$$\underbrace{\frac{1}{3}\left[n+2N\right]\log\frac{n}{\delta} + n}_{D} = \underbrace{\frac{J^*}{C}n^{\frac{4}{3}}\sqrt{\log\frac{n}{\delta}} + N}_{E}.$$

Now write, without loss of generality, $n = \left(A_N N\sqrt{\log N}\right)^{\frac{3}{4}}$, and so

$$D = \frac{1}{3}\left[A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}} + 2N\right]\left[\underbrace{\frac{3}{4}\log A_N + \frac{3}{4}\log N + \frac{3}{8}\log\log N}_{\log n} + \log\frac{1}{\delta}\right] + A_N^{\frac{3}{4}}N^{\frac{3}{4}}(\log N)^{\frac{3}{8}}$$

$$E = \frac{J^*}{C}A_N N\sqrt{\log N}\sqrt{\underbrace{\frac{3}{4}\log A_N + \frac{3}{4}\log N + \frac{3}{8}\log\log N}_{\log n} + \log\frac{1}{\delta}} + N.$$

We will show that $D - E \to 0$ requires $A_N \to C/(\sqrt{3}J^*)$, implying the result.

We first suppose $A_N = \omega(1)$, further breaking into cases which result in different terms inside D and E becoming dominant:

$$\text{If } A_N = \Omega(N), \qquad D = \Theta\left(A_N^{\frac{3}{4}} N^{\frac{3}{4}} (\log N)^{\frac{3}{8}} \log A_N\right), \quad E = \Theta\left(A_N N \sqrt{\log(N)\log(A_N)}\right).$$

$$\text{If } A_N = \Omega\left(\frac{N^{\frac{1}{3}}}{\sqrt{\log N}}\right), A_N = o(N), \quad D = \Theta\left(A_N^{\frac{3}{4}} N^{\frac{3}{4}} (\log N)^{\frac{3}{8}} \log N\right), \quad E = \Theta\left(A_N N \log N\right).$$

$$\text{If } A_N = \omega(1), A_N = o\left(\frac{N^{\frac{1}{3}}}{\sqrt{\log N}}\right), \quad D = \Theta\left(N \log N\right), \qquad E = \Theta\left(A_N N \log N\right).$$

In each case, $E = \omega(D)$ and so $D - E \to -\infty$, contradicting that $D = E$. Thus a stationary point requires $A_N = \mathcal{O}(1)$ for a stationary point.

We now do the same for $A_N = o(1)$. First, clearly $n \geq 1$; suppose that in fact $n = \Theta(1)$, i.e. $A_N = \Theta\left(1/(N\sqrt{\log N})\right)$. In this case, we would have $D = \frac{2}{3} N \log \frac{n}{\delta} + \Theta(1)$ and $E = N + \Theta(1)$, so that $D = E$ requires $\frac{2}{3} \log \frac{n}{\delta} \to 1$, i.e. $n \to \delta \exp \frac{3}{2} \approx 4.5\,\delta$. For $\delta < 0.22$, this contradicts $n \geq 1$. So we know that $\log n = \omega(1)$. Now, the remaining options for A_N all yield $D - E \to \infty$:

$$\text{If } A_N = o(1), A_N = \Omega\left(\frac{1}{\log N}\right), \qquad D = \Theta\left(N \log n\right), \quad E = \Theta\left(A_N N \log n\right).$$

$$\text{If } A_N = o\left(\frac{1}{\log N}\right), A_N = \omega\left(\frac{1}{N\sqrt{\log N}}\right), \quad D = \Theta\left(N \log n\right), \quad E = \Theta\left(N\right).$$

Thus we have established that $A_N = \Theta(1)$. Thus, we obtain that

$$D = \frac{1}{2} N \log N + \mathcal{O}\left(N\right) \qquad E = \frac{\sqrt{3} J^*}{2C} A_N N \log N + \mathcal{O}\left(N\sqrt{\log N}\right).$$

Asymptotic equality hence requires $A_N \to C/(\sqrt{3}J^*)$. □

A.3. Uniform convergence results

These results, on the uniform convergence of $\hat{\eta}$ and $\hat{\sigma}^2$, were used in the proof of Theorem 11.

Proposition 15. *Under Assumptions (A) to (C), we have that with probability at least* $1 - \delta$,

$$\sup_{\omega} |\hat{\eta}_\omega - \eta_\omega| \leq \frac{8}{\sqrt{n}} \left[\nu \sqrt{2 \log \frac{2}{\delta} + 2D \log\left(4R_\Omega \sqrt{n}\right)} + L_k\right].$$

Proof. Theorem 7 of Sriperumbudur et al. (2009) gives a similar bound in terms of Rademacher chaos complexity, but for ease of combination with our bound on convergence of the variance estimator, we use a simple ϵ-net argument instead.

We study the random error function

$$\Delta(\omega) := \hat{\eta}_\omega - \eta_\omega.$$

First, we place T points $\{\omega_i\}_{i=1}^T$ such that for any point $\omega \in \Omega$, $\min_i \|\omega - \omega_i\| \leq q$; Assumption (B) ensures this is possible with at most $T = (4R_\Omega/q)^D$ points (Cucker & Smale, 2001, Proposition 5).

Now, $\mathbb{E}\,\Delta = 0$, because $\hat{\eta}$ is unbiased. Recall that $\hat{\eta} = \frac{1}{n(n-1)} \sum_{i \neq j} H_{ij}$, and via Assumption (A) we know $|H_{ij}| \leq 4\nu$. This $\hat{\eta}$, and hence Δ, satisfies bounded differences: if we replace (X_1, Y_1) with (X_1', Y_1'), obtaining $\hat{\eta}' = \frac{1}{n(n-1)} \sum_{i \neq j} F_{ij}$ where F agrees with H except when i or j is 1, then

$$\begin{aligned}
|\hat{\eta} - \hat{\eta}'| &\leq \frac{1}{n(n-1)} \sum_{i \neq j} |H_{ij} - F_{ij}| = \frac{1}{n(n-1)} \sum_{i > 1} |H_{i1} - F_{i1}| + \frac{1}{n(n-1)} \sum_{j > 1} |H_{1j} - F_{1j}| \\
&\leq \frac{2}{n(n-1)} \sum_{i > 1} 8\nu = \frac{16\nu}{n}.
\end{aligned}$$

Using McDiarmid's inequality for each $\Delta(\omega_i)$ and a union bound, we then obtain that with probability at least $1-\delta$,

$$\max_{i\in\{1,\dots,T\}}|\Delta(\omega)| \leq \frac{16\nu}{\sqrt{2n}}\sqrt{\log\frac{2T}{\delta}} \leq \frac{8\nu}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta}+2D\log\frac{4R_\Omega}{q}}.$$

We also have via Assumption **(C)**, for any two $\omega,\omega'\in\Omega$,

$$|\hat{\eta}_\omega-\hat{\eta}_{\omega'}| \leq \frac{1}{n(n-1)}\sum_{i\neq j}|H_{ij}^{(\omega)}-H_{ij}^{(\omega')}| \leq \frac{1}{n(n-1)}\sum_{i\neq j}4L_k\|\omega-\omega'\| = 4L_k\|\omega-\omega'\|$$
$$|\eta_\omega-\eta_{\omega'}| = |\mathbb{E}\left[H_{12}^{(\omega)}\right]-\mathbb{E}\left[H_{12}^{(\omega')}\right]| \leq \mathbb{E}|H_{12}^{(\omega)}-H_{12}^{(\omega')}| \leq 4L_k\|\omega-\omega'\|$$

so that $\|\Delta\|_L \leq 8L_k$. Combining these two results, we know that with probability at least $1-\delta$

$$\sup_\omega|\Delta(\omega)| \leq \max_{i\in\{1,\dots,T\}}|\Delta(\omega_i)|+8L_kq \leq \frac{8\nu}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta}+2D\log\frac{4R_\Omega}{q}}+8L_kq;$$

setting $q=1/\sqrt{n}$ yields the desired result. □

Proposition 16. *Under Assumptions **(A)** to **(C)**, with probability at least $1-\delta$,*

$$\sup_{\omega\in\Omega}\left|\hat{\sigma}_\omega^2-\sigma_\omega^2\right| \leq \frac{64}{\sqrt{n}}\left[7\sqrt{2\log\frac{2}{\delta}+2D\log\left(4R_\Omega\sqrt{n}\right)}+\frac{18\nu^2}{\sqrt{n}}+8L_k\nu\right].$$

Proof. We again use an ϵ-net argument on the (random) error function

$$\Delta(\omega) := \hat{\sigma}_{k_\omega}^2-\sigma_{k_\omega}^2.$$

First, choose T points $\{\omega_i\}_{i=1}^T$ such that for any point $\omega\in\Omega$, $\min_i\|\omega-\omega_i\|\leq q$; again, via Assumption **(B)** and Proposition 5 of Cucker & Smale (2001) we have $T\leq(4R_\Omega/q)^D$. By Lemmas 17 and 18 and a union bound, with probability at least $1-\delta$,

$$\max_{i\in\{1,\dots,T\}}|\Delta(\omega)| \leq 448\sqrt{\frac{2}{n}\log\frac{2T}{\delta}}+\frac{1152\nu^2}{n} \leq 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}+\frac{2}{n}D\log\frac{4R_\Omega}{q}}+\frac{1152\nu^2}{n}.$$

Lemma 19 shows that $\|\Delta\|_L\leq 512L_k\nu$, which means that with probability at least $1-\delta$,

$$\sup_{\omega\in\Omega}|\Delta(\omega)| \leq 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}+\frac{2}{n}D\log\frac{4R_\Omega}{q}}+\frac{1152\nu^2}{n}+512L_k\nu q. \tag{9}$$

Taking $q=1/\sqrt{n}$ gives the desired result. □

Lemma 17. *For any kernel k bounded by ν (Assumption **(A)**), with probability at least $1-\delta$,*

$$|\hat{\sigma}_k^2-\mathbb{E}\,\hat{\sigma}_k^2| \leq 448\sqrt{\frac{2}{n}\log\frac{2}{\delta}}.$$

Proof. We simply apply McDiarmid's inequality to $\hat{\sigma}_k^2$. Suppose we change (X_1,Y_1) to (X_1',Y_1'), giving a new H matrix F which agrees with H on all but the first row and column. Note that $|H_{ij}|\leq 4\nu$, and recall

$$\hat{\sigma}_k^2 = 4\left(\frac{1}{n^3}\sum_i\left(\sum_j H_{ij}\right)^2-\left(\frac{1}{n^2}\sum_{ij}H_{ij}\right)^2\right).$$

The first term in the parentheses of $\hat{\sigma}_k^2$ changes by

$$|\frac{1}{n^3}\sum_i\left(\sum_j H_{ij}\right)^2 - \frac{1}{n^3}\sum_i\left(\sum_j F_{ij}\right)^2| \leq \frac{1}{n^3}\sum_{ij\ell}|H_{ij}H_{i\ell} - F_{ij}F_{i\ell}|.$$

In this sum, if none of i, j, or ℓ are one, the term is zero. The n^2 terms for which $i = 1$ are each upper-bounded by $32\nu^2$, simply bounding each H or F by 4ν. Of the remainder, there are $(n-1)$ terms where $j = \ell = 1$, each $|H_{i1}^2 - F_{i1}^2| \leq 16\nu^2$. We are left with $2(n-1)^2$ terms which have exactly one of j or ℓ equal to 1; the $j = 1$ terms are $|H_{i1}H_{i\ell} - F_{i1}H_{i\ell}| \leq |H_{i1} - F_{i1}||H_{i\ell}| \leq (8\nu)(4\nu)$, so each of these terms is at most $32\nu^2$. The total sum is thus at most

$$\frac{1}{n^3}\left(n^2 32\nu^2 + (n-1)16\nu^2 + 2(n-1)^2 32\nu^2\right) = \left(\frac{6}{n} - \frac{7}{n^2} + \frac{3}{n^3}\right)16\nu^2.$$

The remainder of the change in $\hat{\sigma}_k^2$ can be determined by bounding

$$\begin{aligned}|\sum_{ij} H_{ij} - \sum_{ij} F_{ij}| \leq \sum_{ij}|H_{ij} - F_{ij}| &= \sum_j |H_{1j} - F_{1j}| + \sum_{i>1}|H_{i1} - F_{i1}| \\ &\leq n(8\nu) + (n-1)(8\nu) = (8\nu)(2n-1),\end{aligned}$$

which then gives us

$$\begin{aligned}|\left(\frac{1}{n^2}\sum_{ij} H_{ij}\right)^2 - \left(\frac{1}{n^2}\sum_{ij} F_{ij}\right)^2| &= |\frac{1}{n^2}\sum_{ij} H_{ij} + \frac{1}{n^2}\sum_{ij} F_{ij}||\frac{1}{n^2}\sum_{ij} H_{ij} - \frac{1}{n^2}\sum_{ij} F_{ij}| \\ &\leq (2\cdot 4\nu)\,\frac{2n-1}{n^2}\,(8\nu) = 64\nu^2\left(\frac{2}{n} - \frac{1}{n^2}\right).\end{aligned}$$

Thus

$$|\hat{\sigma}_k^2 - (\hat{\sigma}_k')^2| \leq 4\left[\left(\frac{6}{n} - \frac{7}{n^2} + \frac{3}{n^3}\right)16\nu^2 + \left(\frac{2}{n} - \frac{1}{n^2}\right)64\nu^2\right] = \frac{64\nu^2}{n^3}\left[14n^2 - 11n + 3\right] \leq \frac{896\nu^2}{n}.$$

Because the same holds for changing any of the (X_i, Y_i) pairs, the result follows by McDiarmid's inequality. □

Lemma 18. *For any kernel k bounded by ν (Assumption (A)), the estimator $\hat{\sigma}_k^2$ satisfies*

$$|\mathbb{E}\,\hat{\sigma}_k^2 - \sigma_k^2| \leq \frac{1152\nu^2}{n}.$$

Proof. We have that

$$\mathbb{E}\,\hat{\sigma}_k^2 = 4\left(\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}\left[H_{i\ell}H_{j\ell}\right] - \frac{1}{n^4}\sum_{ijab}\mathbb{E}\left[H_{ij}H_{ab}\right]\right).$$

Most terms in these sums have their indices distinct; these are the ones that we care about. (We could evaluate the expectations of the other terms exactly, but it would be tedious.) We can thus break down the first term as

$$\begin{aligned}\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}[H_{i\ell}H_{j\ell}] &= \frac{1}{n^3}\sum_{ij\ell:|\{i,j,\ell\}|=3}\mathbb{E}[H_{i\ell}H_{j\ell}] + \frac{1}{n^3}\sum_{ij\ell:|\{i,j,\ell\}|<3}\mathbb{E}[H_{i\ell}H_{j\ell}] \\ &= \frac{n(n-1)(n-2)}{n^3}\,\mathbb{E}[H_{12}H_{13}] + \left(1 - \frac{n(n-1)(n-2)}{n^3}\right)q,\end{aligned}$$

where q is the appropriately-weighted mean of the various $\mathbb{E}[H_{i\ell}H_{j\ell}]$ terms for which i, j, ℓ are not mutually distinct. Since $|H_{ij}| \leq 4\nu$, $\mathbb{E}[H_{i\ell}H_{j\ell}] < 16\nu^2$ and so $|q| \leq 16\nu^2$ as well. Noting that

$$\frac{n(n-1)(n-2)}{n^3} = 1 - \frac{3}{n} + \frac{2}{n^2}$$

we obtain

$$|\frac{1}{n^3}\sum_{ij\ell}\mathbb{E}[H_{i\ell}H_{j\ell}] - \mathbb{E}[H_{12}H_{13}]| = \left(\frac{3}{n} - \frac{2}{n^2}\right)|-\mathbb{E}[H_{12}H_{13}] + q| \leq \left(\frac{3}{n} - \frac{2}{n^2}\right)32\nu^2. \tag{10}$$

The second term can be handled similarly:

$$\begin{aligned}\frac{1}{n^4}\sum_{ijab}\mathbb{E}[H_{ij}H_{ab}] &= \frac{1}{n^4}\sum_{ijab:|\{i,j,a,b\}|=4}\mathbb{E}[H_{ij}H_{ab}] + \frac{1}{n^4}\sum_{ijab:|\{i,j,a,b\}|<4}\mathbb{E}[H_{ij}H_{ab}] \\ &= \frac{n(n-1)(n-2)(n-3)}{n^4}\mathbb{E}[H_{ij}H_{ab}] + \left(1 - \frac{n(n-1)(n-2)(n-3)}{n^4}\right)r,\end{aligned}$$

where r is the appropriately-weighted mean of the non-distinct terms, $|r| \leq 16\nu^2$. For i, j, a, b all distinct, $\mathbb{E}[H_{ij}H_{ab}] = \mathbb{E}[H_{12}]^2$. Here

$$\frac{n(n-1)(n-2)(n-3)}{n^4} = \frac{(n-1)(n^2-5n+6)}{n^3} = 1 - \frac{6}{n} + \frac{11}{n} - \frac{6}{n^3}$$

and so

$$|\frac{1}{n^4}\sum_{ijab}\mathbb{E}[H_{ij}H_{ab}] - \mathbb{E}[H_{12}]^2| \leq \left(\frac{6}{n} - \frac{11}{n^2} + \frac{6}{n^3}\right)32\nu^2. \tag{11}$$

Recalling $\sigma_k^2 = 4(\mathbb{E}[H_{12}H_{13}] - \mathbb{E}[H_{12}]^2)$,

$$|\mathbb{E}\,\hat{\sigma}_k^2 - \sigma_k^2| \leq 128\nu^2\left(\frac{9}{n} - \frac{13}{n^2} + \frac{6}{n^3}\right),$$

and since $n \geq 1$, we have $13/n^2 > 6/n^3$, yielding the result. □

Lemma 19. *Under Assumptions (A) and (C), we have*

$$\sup_{\omega,\omega'\in\Omega}\frac{|\hat{\sigma}_\omega^2 - \hat{\sigma}_{\omega'}^2|}{\|\omega - \omega'\|} \leq 256L_k\nu \qquad \textit{and} \qquad \sup_{\omega,\omega'\in\Omega}\frac{|\sigma_\omega^2 - \sigma_{\omega'}^2|}{\|\omega - \omega'\|} \leq 256L_k\nu.$$

Proof. We first handle the change in $\hat{\sigma}_k$:

$$\begin{aligned}|\hat{\sigma}_{k_\omega}^2 - \hat{\sigma}_{k_{\omega'}}^2| &= 4|\frac{1}{n^3}\sum_{ij\ell}H_{i\ell}^{(\omega)}H_{j\ell}^{(\omega)} - \frac{1}{n^3}\sum_{ij\ell}H_{i\ell}^{(\omega')}H_{j\ell}^{(\omega')} - \frac{1}{n^4}\sum_{ijab}H_{ij}^{(\omega)}H_{ab}^{(\omega)} + \frac{1}{n^4}\sum_{ijab}H_{ij}^{(\omega')}H_{ab}^{(\omega')}| \\ &\leq \frac{4}{n^3}\sum_{ij\ell}|H_{i\ell}^{(\omega)}H_{j\ell}^{(\omega)} - H_{i\ell}^{(\omega')}H_{j\ell}^{(\omega')}| + \frac{4}{n^4}\sum_{ijab}|H_{ij}^{(\omega)}H_{ab}^{(\omega)} - H_{ij}^{(\omega')}H_{ab}^{(\omega')}|.\end{aligned}$$

We can handle both terms by bounding

$$\begin{aligned}|H_{ij}^{(\omega)}H_{ab}^{(\omega)} - H_{ij}^{(\omega')}H_{ab}^{(\omega')}| &\leq |H_{ij}^{(\omega)}H_{ab}^{(\omega)} - H_{ij}^{(\omega)}H_{ab}^{(\omega')}| + |H_{ij}^{(\omega)}H_{ab}^{(\omega')} - H_{ij}^{(\omega')}H_{ab}^{(\omega')}| \\ &= |H_{ij}^{(\omega)}||H_{ab}^{(\omega)} - H_{ab}^{(\omega')}| + |H_{ij}^{(\omega)} - H_{ij}^{(\omega')}||H_{ab}^{(\omega')}| \\ &\leq 4\nu\left(|H_{ab}^{(\omega)} - H_{ab}^{(\omega')}| + |H_{ij}^{(\omega)} - H_{ij}^{(\omega')}|\right).\end{aligned}$$

Using Assumption (C) and the definition of H,

$$|H_{ij}^{(\omega)} - H_{ij}^{(\omega')}| \leq 4L_k\|\omega - \omega'\|$$

so

$$|H_{ij}^{(\omega)}H_{ab}^{(\omega)} - H_{ij}^{(\omega')}H_{ab}^{(\omega')}| \leq 32\nu L_k\|\omega - \omega'\| \tag{12}$$

and hence

$$|\hat{\sigma}_\omega^2 - \hat{\sigma}_{\omega'}^2| \leq 256\nu L_k\|\omega - \omega'\|.$$

Again using (12), we also have

$$\begin{aligned}|\sigma_\omega^2 - \sigma_{\omega'}^2| &\leq 4|\mathbb{E}\left[H_{12}^{(\omega)}H_{13}^{(\omega)}\right] - \mathbb{E}\left[H_{12}^{(\omega')}H_{13}^{(\omega')}\right]| + 4|\mathbb{E}\left[H_{12}^{(\omega)}\right]^2 - \mathbb{E}\left[H_{12}^{(\omega')}\right]^2| \\ &\leq 4\,\mathbb{E}|H_{12}^{(\omega)}H_{13}^{(\omega)} - H_{12}^{(\omega')}H_{13}^{(\omega')}| + 4\,\mathbb{E}|H_{12}^{(\omega)}H_{34}^{(\omega)} - H_{12}^{(\omega')}H_{34}^{(\omega')}| \\ &\leq 256\nu L_k\|\omega - \omega'\|.\end{aligned}$$

□

A.4. Constructing appropriate kernels

We now show Propositions 7 to 9, which each state that Assumption **(C)** is satisfied by various choices of kernel. The following assumption will be useful for different kernel schemes.

(I) The domain $\mathcal{X}$ is Euclidean and bounded, $\mathcal{X} \subseteq \left\{x \in \mathbb{R}^d : \|x\| \leq R_X\right\}$ for some constant $R_X < \infty$.

We begin by recalling a well-known property of the Gaussian kernel, useful for both Gaussian bandwidth selection and deep kernels. A proof is in Appendix A.5.

Lemma 20. *The Gaussian kernel* $\kappa(a,b) = \exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right)$ *satisfies*

$$|\kappa(a,b) - \kappa(a',b')| \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-b\| + \|a'-b'\|\right) \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-a'\| + \|b-b'\|\right).$$

A.4.1. Gaussian bandwidth selection (Proposition 7)

Lemma 20 immediately gives us Assumption **(C)** when we chose among Gaussian kernels:

Proposition 21. *Define a one-dimensional Banach space for inverse lengthscales of Gaussian kernels* $\gamma > 0$*, so that* $k_\gamma(x,y) = \kappa_{1/\gamma}(x,y)$*, with standard addition and multiplication and norms defined by the absolute value, and* k_0 *taken to be the constant* 1 *function. Let* Ω *be any subset of this space. Under Assumption* ***(I)****, Assumption* ***(C)*** *holds: for any* $x, y \in \mathcal{X}$ *and* $\gamma, \gamma' \in \Gamma$*,*

$$|k_\gamma(x,y) - k_{\gamma'}(x,y)| \leq \frac{2R_X}{\sqrt{e}}|\gamma - \gamma'|.$$

Proof.

$$|k_\gamma(x,y) - k_{\gamma'}(x,y)| = |\kappa_1(\gamma x, \gamma y) - \kappa_1(\gamma' x, \gamma' y)| \leq \frac{1}{\sqrt{e}}|\gamma\|x-y\| - \gamma'\|x-y\|| = \frac{\|x-y\|}{\sqrt{e}}|\gamma - \gamma'|.$$

□

A.4.2. Deep kernels (Proposition 9)

To handle the deep kernel case, we will need some more assumptions on the form of the kernel.

(II) $\phi_\omega(x) = \phi_\omega^{(\Lambda)}$ is a feedforward neural network with Λ layers given by

$$\phi_\omega^{(0)}(x) = x \qquad \phi_\omega^{(\ell)}(x) = \sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right),$$

where the network parameter ω consists of all the weight matrices $W_\omega^{(\ell)}$ and biases $b_\omega^{(\ell)}$, and the activation functions $\sigma^{(\ell)}$ are each 1-Lipschitz, $\|\sigma^{(\ell)}(x) - \sigma^{(\ell)}(y)\| \leq \|x-y\|$, with $\sigma^{(\ell)}(0) = 0$ so that $\|\sigma^{(\ell)}(x)\| \leq \|x\|$. Define a Banach space on ω, with addition and scalar multiplication componentwise, and

$$\|\omega\| = \max_{\ell \in \{1,\ldots,\Lambda\}} \max\left(\|W_\omega^{(\ell)}\|, \|b_\omega^{(\ell)}\|\right),$$

where the matrix norm denotes operator norm $\|W\| = \sup_x \|Wx\|/\|x\|$. (For convolutional networks, see Remark 25.)

(III) k_ω is a kernel of the form (1),

$$k_\omega(x,y) = \left[(1-\epsilon)\kappa(\phi_\omega(x),\phi_\omega(y)) + \epsilon\right] q(x,y),$$

with $0 \leq \epsilon \leq 1$, κ a kernel function, and $q(x,y)$ a kernel with $\sup_x q(x,x) \leq Q$.

Note that this includes kernels of the form $k_\omega(x,y) = \kappa(\phi_\omega(x),\phi_\omega(y))$: take $\epsilon = 0$ and $q(x,y) = 1$.

(IV) κ in Assumption (III) is a kernel function satisfying

$$|\kappa(a,b) - \kappa(a',b')| \leq L_\kappa \left(\|a-a'\| + \|b-b'\|\right).$$

This holds for a Gaussian κ via Lemma 20.

We now turn to proving Assumption (C) for deep kernels. First, we will need some smoothness properties of the network ϕ.

Lemma 22. *Under Assumption (II), suppose ω, ω' have $\|\omega\| \leq R$, $\|\omega'\| \leq R$, with $R \neq 1$. Then, for any x,*

$$\|\phi_\omega(x)\| \leq R^\Lambda \|x\| + \frac{R}{R-1}(R^\Lambda - 1) \tag{13}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \left(\Lambda R^{\Lambda-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{R^\Lambda - 1}{(R-1)^2}\right)\|\omega - \omega'\|. \tag{14}$$

If $R \geq 2$, we furthermore have

$$\|\phi_\omega(x)\| \leq R^\Lambda(\|x\| + 2) \tag{15}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \Lambda R^{\Lambda-1}\left(\|x\| + 2\right)\|\omega - \omega'\|. \tag{16}$$

The proof, by recursion, is given in Appendix A.5. We are now ready to prove Assumption (C) for deep kernels.

Proposition 23. *Make Assumptions (I) to (IV) and Assumption (B), with $R_\Omega \geq 2$.*[8] *Then Assumption (C) holds: for any $x, y \in \mathcal{X}$ and $\omega, \omega' \in \Omega$,*

$$|k_\omega(x,y) - k_{\omega'}(x,y)| \leq 2Q(1-\epsilon)L_\kappa \Lambda R_\Omega^{\Lambda-1}(R_X + 2)\|\omega - \omega'\|.$$

Proof.

$$\begin{aligned}
|k_\omega(x,y) - k_{\omega'}(x,y)| &= (1-\epsilon)|\kappa(\phi_\omega(x),\phi_\omega(y)) - \kappa(\phi_{\omega'}(x),\phi_{\omega'}(y))|q(x,y) \\
&\leq Q(1-\epsilon)L_\kappa\left(|\phi_\omega(x) - \phi_{\omega'}(x)| + |\phi_\omega(y) - \phi_{\omega'}(y)|\right) \\
&\leq Q(1-\epsilon)L_\kappa\Lambda R_\Omega^{\Lambda-1}(\|x\| + \|y\| + 4)\|\omega - \omega'\| \\
&\leq Q(1-\epsilon)L_\kappa\Lambda R_\Omega^{\Lambda-1}(2R_X + 4)\|\omega - \omega'\|.
\end{aligned}$$

□

Remark 24. *For the deep kernels we use in the paper (Assumptions (II) to (IV)) on bounded domains (Assumption (I)), we know L_k via Proposition 23; Theorem 6 combines Theorem 11, Corollary 12, and Proposition 23. If we further use a Gaussian kernel q of bandwidth σ_ϕ, the last bracketed term in the error bound of Theorem 11 becomes*

$$\frac{2(1-\epsilon)}{\sigma_\phi\sqrt{e}}\Lambda R_\Omega^{\Lambda-1}(R_X+2) + \sqrt{2\log\frac{2}{\delta} + 2D\log\left(4R_\Omega\sqrt{n}\right)}.$$

The component $R_\Omega^{\Lambda-1}(R_X+2)$, from (15), is approximately the largest that ϕ_ω could make its outputs' norms; σ_ϕ will generally be on a comparable scale to the norm of the actual outputs of the network, so their ratio is something like the "unused capacity" of the network to blow up its inputs. This term is weighted about equally in the convergence bound with the square root of the total number of parameters in the network.

Remark 25. *We can handle convolutional networks as follows. We define Ω in essentially the same way, letting $W_\omega^{(\ell)}$ denote the convolutional kernel (the set of parameters being optimized), but define $\|\omega\|$ in terms of the operator norm of the linear transform corresponding to the convolution operator. This is given in terms of the operator norm of various discrete Fourier transforms of the kernel matrix by Lemma 2 of Bibi et al. (2019); see also Theorem 6 of Sedghi et al. (2019). The number of parameters D is then the actual number of parameters optimized in gradient descent, but the radius R_Ω is computed differently.*

[8] Of course, if we know a bound of $R_\Omega < 2$, the result will still hold using $R_\Omega = 2$. It is also possible to show a tighter result, via (13) and (14) or their analogue for $R = 1$; the expression is simply less compact.

A.4.3. MULTIPLE KERNEL LEARNING (PROPOSITION 8)

Multiple kernel learning (Gönen & Alpaydın, 2011) also falls into our setting. A special case of this family of kernels was studied for the (easier to analyze) "streaming" MMD estimator by Gretton et al. (2012b).

(V) Let $\{k_i\}_{i=1}^D$ be a set of base kernels, each satisfying $\sup_{x\in\mathcal{X}} k_i(x,x) \leq K$ for some finite K. Define k_ω as

$$k_\omega(x,y) = \sum_{i=1}^D \omega_i k_i(x,y).$$

Define the norm of a kernel parameter by the norm of the corresponding vector $\omega \in \mathbb{R}^D$. Let Ω be a set of possible parameters such that for each $\omega \in \Omega$, k_ω is positive semi-definite, and $\|\omega\| \leq R_\Omega$ for some $R_\Omega < \infty$.

Not only does learning in this setting work (Proposition 26), it is also – unlike the deep setting – efficient to find an exact maximizer of $\hat{J}_\lambda$ (Proposition 27).

Proposition 26. *Assumption (V) implies Assumptions (A) to (C). In particular,*

$$\sup_{\omega\in\Omega}\sup_{x\in\mathcal{X}} k_\omega(x,x) \leq K R_\Omega \sqrt{D}$$
$$|k_\omega(x,y) - k_{\omega'}(x,y)| \leq K\sqrt{D}\|\omega - \omega'\|.$$

Proof. Assumption (B) is immediate from Assumption (V), since $\Omega \subset \mathbb{R}^D$. Let $\mathbf{k}(x,y) \in \mathbb{R}^D$ denote the vector whose ith entry is $k_i(x,y)$, so that $k_\omega(x,y) = \omega^\mathsf{T}\mathbf{k}(x,y)$. As $\|\mathbf{k}(x,y)\|_\infty \leq K$, we know $\|\mathbf{k}(x,y)\| \leq K\sqrt{D}$. Assumptions (A) and (C) follow by Cauchy-Schwartz. □

Proposition 27. *Take Assumption (V), and additionally assume that $\Omega = \{\omega \mid \forall i.\, \omega_i \geq 0, \sum_i \omega_i = Q\}$ for some $Q < \infty$. A maximizer of $\hat{J}_\lambda(\omega)$ can then be found by scaling the solution to a convex quadratic program,*

$$\tilde{\omega} = \operatorname*{arg\,min}_{\omega\in[0,\infty)^D\,:\,\omega^\mathsf{T}\mathbf{b}=1} \omega^\mathsf{T}(\mathbf{A}+\lambda I)\omega, \qquad \hat{\omega} = \frac{Q}{\sum_i \tilde{\omega}_i}\tilde{\omega} \in \operatorname*{arg\,max}_{\omega\in\Omega} \hat{J}_\lambda(\omega),$$

where

$$(\mathbf{H}_{ij})_\ell = k_\ell(X_i,X_j) + k_\ell(Y_i,Y_j) - k_\ell(X_i,Y_j) - k_\ell(X_j,Y_i)$$
$$\mathbf{b} = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{H}_{ij} \in \mathbb{R}^D$$
$$\mathbf{A} = \frac{4}{n^3}\sum_i\left(\sum_j \mathbf{H}_{ij}\right)\left(\sum_j \mathbf{H}_{ij}\right)^\mathsf{T} - \frac{4}{n^4}\left(\sum_{ij}\mathbf{H}_{ij}\right)\left(\sum_{ij}\mathbf{H}_{ij}\right)^\mathsf{T} \in \mathbb{R}^{D\times D},$$

as long as $\mathbf{b}$ *has at least one positive entry.*

Proof. The H matrix used by $\hat{\eta}_\omega$ and $\hat{\sigma}_\omega$ takes a simple form:

$$H_{ij}^{(\omega)} = k_\omega(X_i,X_j) + k_\omega(Y_i,Y_j) - k_\omega(X_i,Y_j) - k_\omega(X_j,Y_i) = \omega^\mathsf{T}\mathbf{H}_{ij}.$$

Thus

$$\hat{\eta}_\omega = \omega^\mathsf{T}\left(\frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{H}_{ij}\right) = \omega^\mathsf{T}\mathbf{b}$$
$$\hat{\sigma}_\omega^2 = \frac{4}{n^3}\sum_i\left(\omega^\mathsf{T}\sum_j\mathbf{H}_{ij}\right)^2 - \frac{4}{n^4}\left(\omega^\mathsf{T}\sum_{ij}\mathbf{H}_{ij}\right)^2$$
$$= \omega^\mathsf{T}\left(\frac{4}{n^3}\sum_i\left(\sum_j\mathbf{H}_{ij}\right)\left(\sum_j\mathbf{H}_{ij}\right)^\mathsf{T} - \frac{4}{n^4}\left(\sum_{ij}\mathbf{H}_{ij}\right)\left(\sum_{ij}\mathbf{H}_{ij}\right)^\mathsf{T}\right)\omega = \omega^\mathsf{T}\mathbf{A}\omega.$$

Note that because $\hat{\sigma}_\omega^2 \geq 0$ for any ω, we have $\mathbf{A} \succeq 0$. We have now obtained a problem equivalent to the one in Section 4 of Gretton et al. (2012b); the argument proceeds as there. □

A.5. Miscellaneous Proofs

The following lemma was used for Propositions 21 and 23.

Lemma 20. *The Gaussian kernel* $\kappa(a,b) = \exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right)$ *satisfies*

$$|\kappa(a,b) - \kappa(a',b')| \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-b\| + \|a'-b'\|\right) \leq \frac{1}{\sigma\sqrt{e}}\left(\|a-a'\| + \|b-b'\|\right).$$

Proof. We have that

$$\begin{aligned}|\kappa(a,b) - \kappa(a',b')| &= |\exp\left(-\frac{\|a-b\|^2}{2\sigma^2}\right) - \exp\left(-\frac{\|a'-b'\|^2}{2\sigma^2}\right)| \\ &\leq \|x \mapsto \exp\left(-\frac{x^2}{2\sigma^2}\right)\|_L \,|\|a-b\| - \|a'-b'\||.\end{aligned}$$

We can bound the Lipschitz constant as its maximal derivative norm,

$$\sup_x \frac{|x|}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Noting that

$$\frac{\mathrm{d}}{\mathrm{d}x}\log\left(\frac{|x|}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\right) = \frac{1}{x} - \frac{x}{\sigma^2}$$

vanishes only at $x = \pm\sigma$, the supremum is achieved by using that value, giving

$$\|x \mapsto \exp\left(-\frac{x^2}{2\sigma^2}\right)\|_L = \frac{1}{\sigma\sqrt{e}}.$$

The result follows from

$$|\|a-b\| - \|a'-b'\|| \leq \|a-b-a'+b'\| \leq \|a-a'\| + \|b-b'\|.$$ □

This next lemma was used in Proposition 23.

Lemma 22. *Under Assumption (II), suppose* ω, ω' *have* $\|\omega\| \leq R$, $\|\omega'\| \leq R$, *with* $R \neq 1$. *Then, for any x,*

$$\|\phi_\omega(x)\| \leq R^\Lambda \|x\| + \frac{R}{R-1}(R^\Lambda - 1) \tag{13}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \left(\Lambda R^{\Lambda-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{R^\Lambda - 1}{(R-1)^2}\right)\|\omega - \omega'\|. \tag{14}$$

If $R \geq 2$, *we furthermore have*

$$\|\phi_\omega(x)\| \leq R^\Lambda(\|x\| + 2) \tag{15}$$

$$\|\phi_\omega(x) - \phi_{\omega'}(x)\| \leq \Lambda R^{\Lambda-1}\left(\|x\| + 2\right)\|\omega - \omega'\|. \tag{16}$$

Proof. First, $\|\phi_\omega^{(0)}(x)\| = \|x\|$, showing (13) when $\Lambda = 0$. In general,

$$\begin{aligned}\|\phi_\omega^{(\ell)}(x)\| &= \|\sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right)\| \\ &\leq \|W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\| \\ &\leq \|W_\omega^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x)\| + \|b_\omega^{(\ell)}\| \\ &\leq R\|\phi_\omega^{(\ell-1)}(x)\| + R,\end{aligned}$$

and expanding this recursion gives

$$\|\phi_\omega^{(\ell)}(x)\| \leq R^\ell \|x\| + \sum_{m=1}^{\ell} R^m = R^\ell \|x\| + \frac{R}{R-1}(R^\ell - 1).$$

Now, we have (14) for $\Lambda = 0$ because $\phi_\omega^{(0)}(x) - \phi_{\omega'}^{(0)}(x) = 0$. For $\ell \geq 1$, we have

$$\begin{aligned}
\|\phi_\omega^{(\ell)}(x) - \phi_{\omega'}^{(\ell)}(x)\| &= \|\sigma^{(\ell)}\left(W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) + b_\omega^{(\ell)}\right) - \sigma^{(\ell)}\left(W_{\omega'}^{(\ell)}\phi_{\omega'}^{(\ell-1)}(x) - b_{\omega'}^{(\ell)}\right)\| \\
&\leq \|W_\omega^{(\ell)}\phi_\omega^{(\ell-1)}(x) - W_{\omega'}^{(\ell)}\phi_\omega^{(\ell-1)}(x)\| + \|W_{\omega'}^{(\ell)}\phi_\omega^{(\ell-1)}(x) - W_{\omega'}^{(\ell)}\phi_{\omega'}^{(\ell-1)}(x)\| + \|b_\omega^{(\ell)} - b_{\omega'}^{(\ell)}\| \\
&\leq \|W_\omega^{(\ell)} - W_{\omega'}^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x)\| + \|W_{\omega'}^{(\ell)}\|\|\phi_\omega^{(\ell-1)}(x) - \phi_{\omega'}^{(\ell-1)}(x)\| + \|\omega - \omega'\| \\
&\leq \|\omega - \omega'\|\left(R^{\ell-1}\|x\| + \frac{R}{R-1}(R^{\ell-1} - 1) + 1\right) + R\|\phi_\omega^{(\ell-1)}(x) - \phi_{\omega'}^{(\ell-1)}(x)\|.
\end{aligned}$$

Expanding the recursion yields

$$\begin{aligned}
\|\phi_\omega^{(\ell)}(x) - \phi_{\omega'}^{(\ell)}(x)\| &\leq \sum_{m=0}^{\ell-1} R^m \left(R^{\ell-1-m}\|x\| + \frac{R}{R-1}(R^{\ell-m-1} - 1) + 1\right)\|\omega - \omega'\| \\
&= \sum_{m=0}^{\ell-1}\left(R^{\ell-1}\|x\| + \frac{R^\ell}{R-1} - \frac{R^{m+1}}{R-1} + R^m\right)\|\omega - \omega'\| \\
&= \left(\ell R^{\ell-1}\|x\| + \frac{\ell R^\ell}{R-1} - \left(\frac{R}{R-1} - 1\right)\sum_{m=0}^{\ell-1} R^m\right)\|\omega - \omega'\| \\
&= \left(\ell R^{\ell-1}\left(\|x\| + \frac{R}{R-1}\right) - \frac{1}{R-1}\frac{R^\ell - 1}{R-1}\right)\|\omega - \omega'\|.
\end{aligned}$$

When $R \geq 2$, we have that $R/(R-1) \leq 2$ and $R^\ell > 1$, giving (15) and (16). □

B. Experimental Details

B.1. Details of synthetic datasets

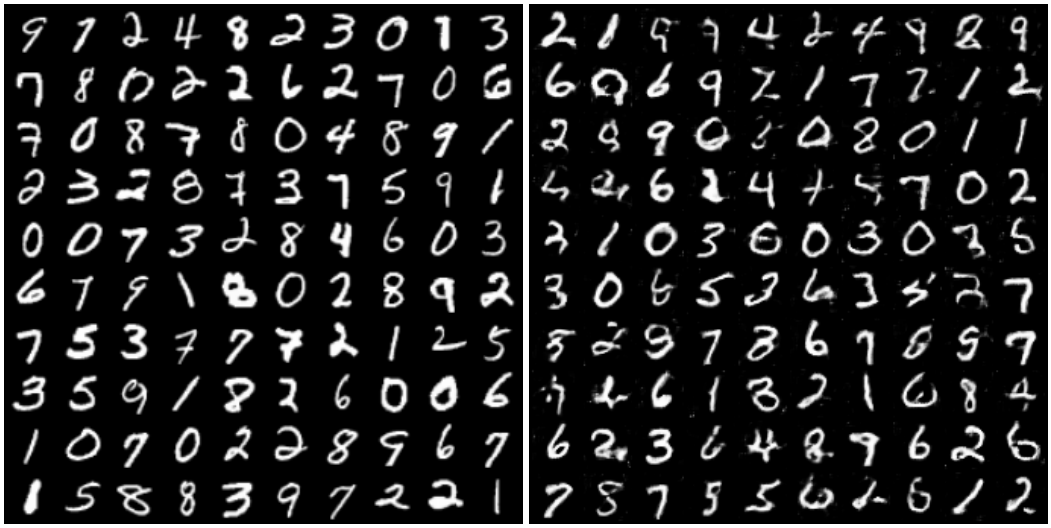
Table 6 shows details of four synthetic datasets. *Blob* datasets are often used to validate two-sample test methods (Gretton et al., 2012b; Jitkrittum et al., 2016; Sutherland et al., 2017), although we rotate each blob to show the benefits of non-homogeneous kernels. *HDGM* datasets are first proposed in this paper. *HDGM-D* can be regarded as *high-dimension Blob-D* which contains two modes with the same variance and different covariance.

Table 6. Specifications of $\mathbb{P}$ and $\mathbb{Q}$ of synthetic datasets. $\mu_1^b = [0,0], \mu_2^b = [0,1], \mu_3^b = [0,2], \ldots, \mu_8^b = [2,1], \mu_9^b = [2,2]$ (same with Figure 1a). $\mu_1^h = \mathbf{0}_d$, $\mu_2^h = 0.5 \times \mathbf{1}_d$, I_d is an identity matrix with size d. $\Delta_i^b = -0.02 - 0.002 \times (i-1)$ if $i < 5$ and $\Delta_i^b = 0.02 + 0.002 \times (i-6)$ if $i > 5$. if $i = 5$, $\Delta_i^b = 0$ (same with Figure 1a). Δ_1^h and Δ_2^h are set to 0.5 and -0.5, respectively.

Datasets	$\mathbb{P}$	$\mathbb{Q}$
Blob-S	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$
Blob-D	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9} \frac{1}{9}\mathcal{N}\left(\mu_i^b, \begin{bmatrix} 0.03 & \Delta_i^b \\ \Delta_i^b & 0.03 \end{bmatrix}\right)$
HDGM-S	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$
HDGM-D	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$	$\sum_{i=1}^{2} \frac{1}{2}\mathcal{N}\left(\mu_i^h, \begin{bmatrix} 1 & \Delta_i^h & \mathbf{0}_{d-2} \\ \Delta_i^h & 1 & \mathbf{0}_{d-2} \\ \mathbf{0}_{d-2}^T & \mathbf{0}_{d-2}^T & I_{d-2} \end{bmatrix}\right)$

B.2. Dataset visualization

Figure 4 shows images from real-*MNIST* and "fake"-*MNIST*, while Figure 5 shows samples from *CIFAR-10* and *CIFAR-10.1*.



(a) Real-*MNIST* (b) "Fake"-*MNIST*

Figure 4. Images from real-*MNIST* and "fake"-*MNIST*. "Fake"-*MNIST* is generated by DCGAN (Radford et al., 2016).

B.3. Configurations

We implement all methods on Python 3.7 (Pytorch 1.1) with a NIVIDIA Titan V GPU. We run ME and SCF using the official code (Jitkrittum et al., 2016), and implement C2ST-S, C2ST-L, MMD-D and MMD-O by ourselves. We use permutation test to compute p-values of C2ST-S and C2ST-L, MMD-D, MMD-O and tests in Table 4. We set $\alpha = 0.05$ for all experiments. Following Lopez-Paz & Oquab (2017), we use a deep neural network F as the classifier in C2ST-S and C2ST-L, and train the F by minimizing cross entropy. To fairly compare MMD-D with C2ST-S and C2ST-L, the network ϕ_ω in MMD-D has the same architecture with feature extractor in F. Namely, $F = g \circ \phi_\omega$, where g is a two-layer fully-connected network. The network g is a simple binary classifier that takes extracted features (through ϕ_ω) as input. For test methods shown in Table 4, the network ϕ_ω in them also has the same architecture with that in MMD-D.

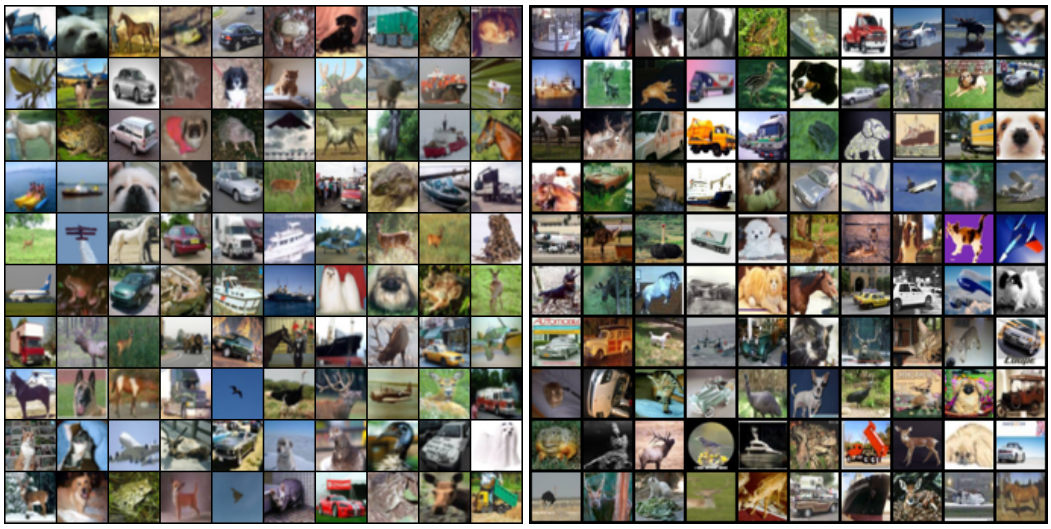
For *Blob*, *HDGM* and *Higgs*, ϕ_ω is a five-layer fully-connected neural network. The number of neurons in hidden and output layers of ϕ_ω are set to 50 for *Blob*, $3 \times d$ for *HDGM* and 20 for *Higgs*, where d is the dimension of samples. These neurons are with softplus activation function, i.e., $\log(1 + \exp(x))$. For *MNIST* and *CIFAR*, ϕ_ω is a *convolutional neural network* (CNN) that contains four convolutional layers and one fully-connected layer. The structure of the CNN follows the structure of the feature extractor in the discriminator of DCGAN (Radford et al., 2016) (see Figures 6 and 8 for the structure of ϕ_ω in MMD-D, and Figures 7 and 9 for the structure of classifier F in C2ST-S and C2ST-L). The link of DCGAN code is `https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/dcgan/dcgan.py`.

We use Adam optimizer (Kingma & Ba, 2015) to optimize 1) parameters of F in C2ST-S and C2ST-L, 2) parameters of ϕ_ω in MMD-D and 3) kernel lengthscale in MMD-O. We set drop-out rate to zero when training C2ST-S, C2ST-L and MMD-D on all datasets.

B.4. Detailed parameters of all test methods

In this subsection, we demonstrate detailed parameters of all test methods. Except for learning rate of Adam optimizer, we use default parameters of Adam optimizer provided by Pytorch. We use one validation set (with the same size of training set) to roughly search these parameters. Using these parameters, we compute test power of each test method on 100 test sets (with the same size of training set).

For ME and SCF, we follow Chwialkowski et al. (2015) and set $J = 10$ for *Higgs*. For other datasets, we set $J = 5$.



(a) *CIFAR-10* test set (b) *CIFAR-10.1* test set

Figure 5. Images from *CIFAR-10* test set and the new *CIFAR-10.1* test set (Recht et al., 2019).

Figure 6. The structure of ϕ_ω in MMD-D on *MNIST*. The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. Best viewed zoomed in.

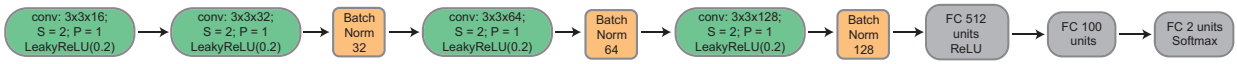


Figure 7. The structure of classifier F in C2ST-S and C2ST-L on *MNIST*. The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. In the first layer, we will convert the *CIFAR* images from $32 \times 32 \times 3$ to $64 \times 64 \times 3$. Best viewed zoomed in.

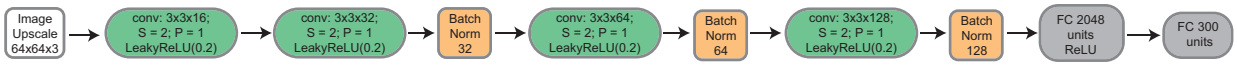


Figure 8. The structure of ϕ_ω in MMD-D on *CIFAR*. The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout in all layers. In the first layer, we will convert the *CIFAR* images from $32 \times 32 \times 3$ to $64 \times 64 \times 3$. Best viewed zoomed in.

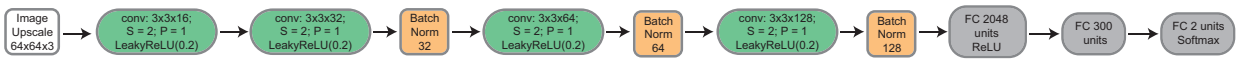


Figure 9. The structure of classifier F in C2ST-S and C2ST-L on *CIFAR*. The kernel size of each convolutional layer is 3; stride (S) is set to 2; padding (P) is set to 1. We do not use dropout. Best viewed zoomed in.

For C2ST-S and C2ST-L, we set batchsize to $\min\{2 \times n_b, 128\}$ for *Blob*, 128 for *HDGM* and *Higgs*, and 100 for *MNIST* and *CIFAR*. We set the number of epochs to $500 \times 18 \times n_b/\text{batchsize}$ for *Blob*, $1,000$ for *HDGM*, *Higgs* and *CIFAR*, and $2,000$ for *MNIST*. We set learning rate to 0.001 for *Blob*, *HDGM* and *Higgs*, and 0.0002 for *MNIST* and *CIFAR* (following Radford et al. (2016)).

For MMD-O, we use full batch (i.e., all samples) to train MMD-O. we set the number of epochs to $1,000$ for *Blob*, *HDGM*, *Higgs* and *CIFAR*, and $2,000$ for *MNIST*. We set learning rate to 0.0005 for *Blob*, *MNIST* and *CIFAR*, and 0.001 for *HDGM*.

Table 7. Results on *Higgs* ($\alpha = 0.05$). We report average Type I error on *Higgs* dataset when increasing number of samples (N). Note that, in *Higgs*, we have two types of Type I errors: 1) Type I error when two samples drawn from $\mathbb{P}$ (no Higgs bosons) and 2) Type I error when two samples drawn from $\mathbb{Q}$ (having Higgs bosons). Type I reported here is the average value of 1) and 2). Since Type I error reported here is the average value of two average Type I errors, we do not report standard errors of the average Type I error in this table.

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
1000	0.048	0.040	0.043	0.048	0.059	0.037
2000	0.043	0.032	0.060	0.056	0.055	0.053
3000	0.049	0.043	0.046	0.053	0.051	0.069
5000	0.056	0.035	0.052	0.065	0.049	0.062
8000	0.050	0.034	0.065	0.067	0.056	0.037
10000	0.059	0.032	0.057	0.058	0.045	0.048
Avg.	0.051	0.036	0.054	0.058	0.050	0.051

Table 8. Results on *MNIST* given $\alpha = 0.05$. We report average Type I error$\pm$standard errors on real-*MNIST* vs. real-*MNIST* when increasing number of samples (N).

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-D
200	0.076$\pm$0.011	0.075$\pm$0.010	0.035$\pm$0.006	0.045$\pm$0.005	0.068$\pm$0.004	0.056$\pm$0.003
400	0.062$\pm$0.010	0.056$\pm$0.007	0.044$\pm$0.006	0.040$\pm$0.004	0.053$\pm$0.005	0.056$\pm$0.005
600	0.051$\pm$0.003	0.049$\pm$0.009	0.039$\pm$0.005	0.054$\pm$0.007	0.066$\pm$0.008	0.056$\pm$0.008
800	0.054$\pm$0.006	0.046$\pm$0.006	0.043$\pm$0.005	0.042$\pm$0.007	0.051$\pm$0.005	0.054$\pm$0.007
1000	0.047$\pm$0.006	0.045$\pm$0.010	0.038$\pm$0.006	0.046$\pm$0.005	0.041$\pm$0.007	0.062$\pm$0.006
Avg.	0.058	0.054	0.040	0.045	0.056	0.057

For MMD-D, we use full batch (i.e., all samples) to train MMD-D with samples from *Blob*, *HDGM* and *Higgs*. We use mini-batch (batchsize is 100) to train MMD-D with samples from *MNIST* and *CIFAR*. We set the number of epochs to $1,000$ for *Blob*, *HDGM*, *Higgs* and *CIFAR*, and $2,000$ for *MNIST*. We set learning rate to 0.0005 for *Blob* and *Higgs*, 10^{-5} for *HDGM*, 0.001 for *MNIST* and 0.0002 for and *CIFAR* (following Radford et al. (2016)).

B.5. Links to datasets

Higgs dataset can be downloaded from UCI Machine Learning Repository. The link is `https://archive.ics.uci.edu/ml/datasets/HIGGS`.

MNIST dataset can be downloaded via Pytorch. See the code in `https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/dcgan/dcgan.py`.

CIFAR-10.1 is available from `https://github.com/modestyachts/CIFAR-10.1/tree/master/datasets` (we use `cifar10.1_v4_data.npy`). This new test set contains $2,031$ images from TinyImages (Torralba et al., 2008).

B.6. Type I errors on *Higgs* and *MNIST*

Table 7 shows average Type I error on *Higgs* dataset when increasing number of samples (N). Table 8 shows average Type I error on real-*MNIST* vs. real-*MNIST* when increasing number of samples (N).

C. Interpretability on *CIFAR-10* vs *CIFAR-10.1*

In Section 7.1, we have shown that images in *CIFAR-10* and *CIFAR-10.1* are not from the same distribution. Thus, it is interesting to try to understand the major difference between the datasets. Mean Embedding tests (Chwialkowski et al., 2015) compare the mean embeddings $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ at test locations $v_1, \dots, v_L$, rather than through their overall norm. The test statistic is

$$\hat{\Lambda} = n\bar{z}_n^{\mathsf{T}} S^{-1} \bar{z}_n, \quad z_i = (k(x_i, v_j) - k(y_i, v_j))_{j=1}^{L} \in \mathbb{R}^L, \quad \bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad S_n = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z}_n)(z_i - \bar{z}_n)^{\mathsf{T}};$$

the asymptotic null distribution of $\hat{\Lambda}$ is χ^2_L, and the estimator is computable in linear time rather than $\widehat{\mathrm{MMD}}_U$'s quadratic time.

Jitkrittum et al. (2017) jointly learn the parameters v_j and kernel parameters to optimize test power. The best such test locations ($L = 1$) for a Gaussian kernel (with learned bandwidth) are shown in Figure 10. We could also try optimizing a deep kernel (1) and the test locations together; this procedure, however, failed to find a useful test. We can find a better test, though, with a two-stage scheme: first, learn a deep kernel to maximize $\hat{J}_\lambda$, then choose v_i to maximize $\hat{\Lambda}$ with that kernel fixed. Results are shown in Figure 11.

Although these approaches give nontrivial test power, it is hard to interpret either set of images, as the test locations have moved far outside the set of natural images. We can instead constrain $v_1 \in S_{\mathbb{P}} \cup S_{\mathbb{Q}}$, simply picking the single point from the dataset which maximizes $\hat{\Lambda}$ (shown in Figure 12). This achieves similar test power, but lets us see that the difference might lie in images with smaller objects of interest than the mean for *CIFAR-10*.

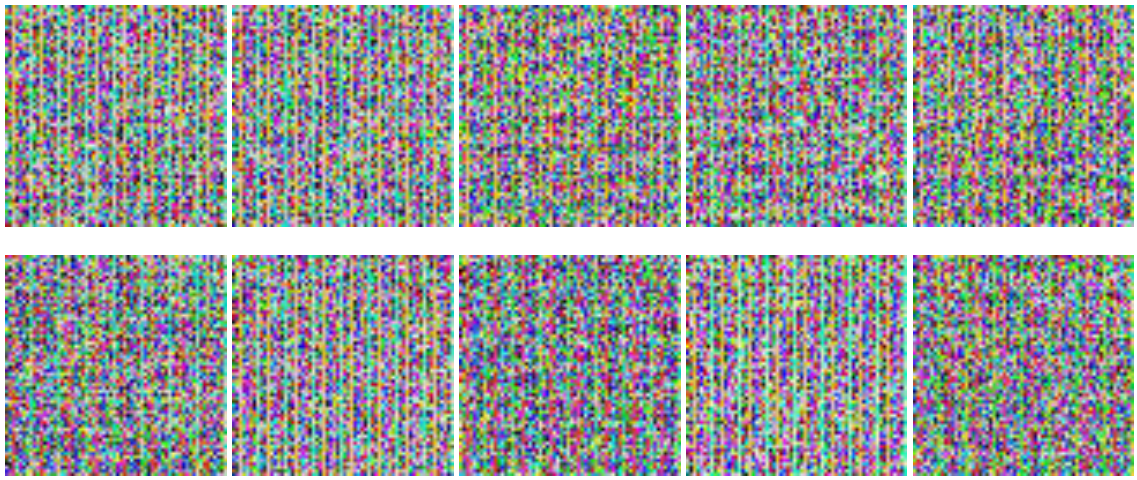


Figure 10. The best test locations (learned by an ME test with $L = 1$) from 10 experiments on *CIFAR-10* vs *CIFAR-10.1*. Average rejection rate is 0.415.

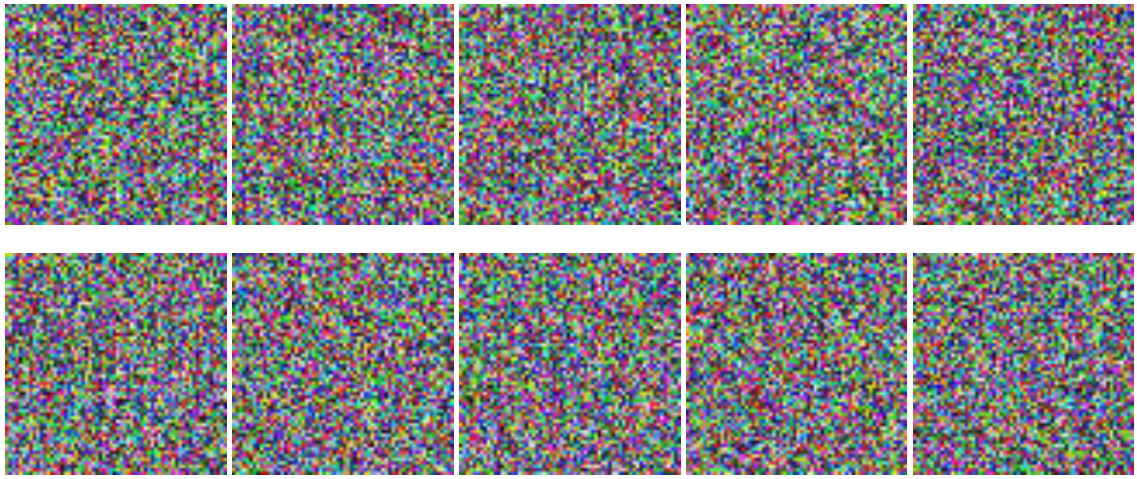


Figure 11. The best test locations (learned by an ME test, $L = 1$, with a deep kernel optimized for an MMD test) from 10 experiments on *CIFAR-10* vs *CIFAR-10.1*. Average rejection rate is 0.637.

Figure 12. The best test locations (selected among existing images with our learned deep kernel, $L = 1$) from 10 experiments on *CIFAR-10* vs *CIFAR-10.1*. Average rejection rate is 0.653.