

## 809 A PROOFS

### 810 A.1 Discriminative properties (proofs of section 3)

#### 811 A.1.1 Comparison with other distances (proofs of subsection 3.1)

812 **Lemma 3.3.** The Schatten 2-norm of the difference of the kernel density operators (using kernel  $k$ )  
813 of two distributions corresponds to their Maximum Mean Discrepancy using the kernel  $k^2$ :

$$\|\Sigma_\mu - \Sigma_\nu\|_2 = MMD_{k^2}(\mu, \nu) \quad (12)$$

814 Consequently,  $MMD_{k^2}(\mu, \nu) \leq d_{KT}(\mu, \nu)$

815 PROOF. We have :

$$\begin{aligned} \langle \Sigma_\mu, \Sigma_\nu \rangle &= Tr(\Sigma_\mu \Sigma_\nu^*) = Tr(\Sigma_\mu \Sigma_\nu) \\ &= Tr\left(\int_{\mathcal{X}} \varphi(x) \varphi(x)^* \mu(x) dx \int_{\mathcal{Y}} \varphi(y) \varphi(y)^* \nu(y) dy\right) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} Tr(\varphi(x) \varphi(x)^* \varphi(y) \varphi(y)^*) \mu(x) \nu(y) dx dy \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} k(x, y) k(x, y) \mu(x) \nu(y) dx dy \end{aligned}$$

816 And  $\|\Sigma_\mu - \Sigma_\nu\|_2^2 = \langle \Sigma_\mu, \Sigma_\mu \rangle + \langle \Sigma_\nu, \Sigma_\nu \rangle - 2\langle \Sigma_\mu, \Sigma_\nu \rangle$ , hence the result.  $\square$

817 **Lemma A.1.** For the Gaussian kernel  $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ , we have:

$$\|\varphi(x) - \varphi(y)\|_{\mathcal{H}} \leq \frac{\|x - y\|}{\sigma}$$

818 PROOF.

$$\begin{aligned} \|\varphi(x) - \varphi(y)\|_{\mathcal{H}} &= \sqrt{\|\varphi(x)\|_{\mathcal{H}}^2 + \|\varphi(y)\|_{\mathcal{H}}^2 - 2k(x, y)} \\ &= \sqrt{2(1 - k(x, y))} \\ &\leq \sqrt{2 \frac{\|x - y\|^2}{2\sigma^2}} \\ &= \frac{\|x - y\|}{\sigma} \end{aligned}$$

819 where we used  $1 - e^{-x} \leq x$ .  $\square$

820 Thanks to Lemma A.1, we can prove Corollary 3.4:

821 **Corollary 3.4.** If Assumption 1 is verified,

$$d_{KT}(\mu, \nu) \leq 2W_{c_k}(\mu, \nu).$$

822 Furthermore, using the Gaussian kernel with parameter  $\sigma$ ,

$$d_{KT}(\mu, \nu) \leq 2W_{c_k}(\mu, \nu) \leq \frac{2}{\sigma} W_{\|\cdot\|}(\mu, \nu).$$

823 PROOF. For any coupling  $\pi$  of  $\mu$  and  $\nu$ , any  $f \in \mathcal{F}_1$ , thanks to Proposition 3.1:

$$\begin{aligned} \mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(Y)] &= \mathbb{E}_\pi[f(X) - f(Y)] \\ &\leq 2\mathbb{E}_\pi\|\varphi(X) - \varphi(Y)\|_{\mathcal{H}} \end{aligned}$$

824 This is true for any coupling and any function of  $\mathcal{F}_1$  so it stays true when minimising over all couplings  
825 and maximising over all  $f \in \mathcal{F}_1$ . The second inequality follows similarly from Lemma A.1

826 Finally, we show how to use the Fuchs and Van De Graaf [1999] inequalities to frame the kernel trace  
827 distance with the Kernel Bures-Wasserstein distance:

828 **Lemma A.2.** When Assumption 1 is verified,

$$d_{KBW}(\mu, \nu)^2 \leq d_{KT}(\mu, \nu) \leq 2d_{KBW}(A, B) \quad (13)$$

829 PROOF.

830 We have  $F(\Sigma_\mu, \Sigma_\nu) = 1 - \frac{1}{2}d_{KBW}(\mu, \nu)^2$ . Therefore, from  $2(1 - F(\Sigma_\mu, \Sigma_\nu)) \leq \|\Sigma_\mu - \Sigma_\nu\|_1$ , we  
 831 get  $d_{KBW}(\mu, \nu)^2 \leq d_{KT}(\mu, \nu)$ . For the second inequality:

$$\begin{aligned} \|\Sigma_\mu - \Sigma_\nu\|_1 &\leq 2\sqrt{1 - F(\Sigma_\mu, \Sigma_\nu)^2} \\ &= 2\sqrt{1 - (1 - \frac{1}{2}d_{KBW}(\mu, \nu)^2)^2} \\ &= 2\sqrt{1 - (1 - \frac{1}{4}d_{KBW}(\mu, \nu)^4 - d_{KBW}(\mu, \nu)^2)} \\ &= 2\sqrt{d_{KBW}(\mu, \nu)^2 - \frac{1}{4}d_{KBW}(\mu, \nu)^4} \\ &= 2d_{KBW}(\mu, \nu)\sqrt{1 - \frac{1}{4}d_{KBW}(\mu, \nu)^2} \\ &\leq 2d_{KBW}(\mu, \nu) \end{aligned}$$

832

□

## 833 A.2 Proof of Proposition 3.1

834 PROOF. The definition as an IPM (i) comes from the dual definition in Eq. (2). Indeed,

$$\|\Sigma_\mu - \Sigma_\nu\|_1 = \sup_{U \in \mathcal{L}(\mathcal{H}), \|U\|_\infty = 1} \langle U, \Sigma_\mu - \Sigma_\nu \rangle.$$

835 Then,  $\langle U, \Sigma_\mu \rangle = \text{Tr}(U^* \mathbb{E}_{X \sim \mu}[\varphi(X)\varphi(X)^*]) = \mathbb{E}_{X \sim \mu}[\text{Tr}(U^* \varphi(X)\varphi(X)^*)]$  and similarly for  $\Sigma_\nu$ .  
 836 For symmetric (around zero) space of functions, one can drop the absolute values in the definition  
 837 of IPM. So the formulation of  $\mathcal{F}_1$  comes from  $\text{Tr}(U^* \varphi(x)\varphi(x)^*) = \varphi(x)^* U^* \varphi(x)$  (equivalently,  
 838 picking  $U^*$  instead of  $U$ , since adjunction applied to  $\{U \in \mathcal{L}(\mathcal{H}), \|U\|_\infty = 1\}$  is a bijection from  
 839 this space to itself, gives  $\varphi(x)^* U \varphi(x)$ ).

840 Regarding (ii), for  $x \in \mathcal{X}$ ,  $U \in \mathcal{L}(\mathcal{H})$  with  $\|U\|_\infty = 1$ :

$$\begin{aligned} |\varphi(x)^* U \varphi(x)| &\leq \|\varphi(x)\|_{\mathcal{H}} \|U \varphi(x)\|_{\mathcal{H}} \\ &\leq \|\varphi(x)\|_{\mathcal{H}} \|\varphi(x)\|_{\mathcal{H}} = 1, \end{aligned}$$

841 where the equality comes from Assumption 1.

842 Then concerning (iii) for  $x, y \in \mathcal{X}$ :

$$\begin{aligned} |f(x) - f(y)| &= |\varphi(x)^* U \varphi(x) - \varphi(y)^* U \varphi(y)| \\ &= |\varphi(x)^* U \varphi(x) - \varphi(y)^* U \varphi(x) + \varphi(y)^* U \varphi(x) - \varphi(y)^* U \varphi(y)| \\ &= |(\varphi(x) - \varphi(y))^* U \varphi(x) - \varphi(y)^* U (\varphi(y) - \varphi(x))| \\ &\leq |(\varphi(x) - \varphi(y))^* U \varphi(x)| + |\varphi(y)^* U (\varphi(y) - \varphi(x))| \\ &\leq \|(\varphi(x) - \varphi(y))\| \cdot \|U \varphi(x)\| + \|\varphi(y)\| \cdot \|U(\varphi(y) - \varphi(x))\| \end{aligned}$$

843 by Cauchy-Schwartz. Because  $\|U\|_\infty \leq 1$ :

$$|f(x) - f(y)| \leq \|\varphi(x) - \varphi(y)\| \cdot \|\varphi(x)\| + \|\varphi(y)\| \cdot \|\varphi(y) - \varphi(x)\|$$

844 therefore by Assumption 1:

$$|f(x) - f(y)| \leq 2\|\varphi(x) - \varphi(y)\|$$

845

□

### 846 A.2.1 Normalised energy (proof of subsection 3.2)

847 **Proposition 3.5.** Let's consider distances between two mixtures  $P = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2$  and  $Q = \frac{1}{2}\nu_1 + \frac{1}{2}\nu_2$   
 848 such that  $\Sigma_{\mu_1}, \Sigma_{\nu_1}$  are orthogonal to  $\Sigma_{\mu_2}, \Sigma_{\nu_2}$ . Then:

$$\begin{aligned} d_{KT}(P, Q) &= \frac{1}{2}d_{KT}(\mu_1, \nu_1) + \frac{1}{2}d_{KT}(\mu_2, \nu_2) \\ MMD_{k^2}^2(P, Q) &= \frac{1}{4}MMD_{k^2}^2(\mu_1, \nu_1) + \frac{1}{4}MMD_{k^2}^2(\mu_2, \nu_2). \end{aligned}$$

PROOF. Noting  $\alpha = \mu_1 - \nu_1$  and  $\beta = \mu_2 - \nu_2$ , notice that  $\Sigma_P - \Sigma_Q = \Sigma_\alpha + \Sigma_\beta$ , and that  $\Sigma_\alpha \perp \Sigma_\beta$ . Looking at the dual expression (2) of the Schatten norm:

$$\|\frac{1}{2}\Sigma_\alpha + \frac{1}{2}\Sigma_\beta\|_1 = \sup_{U \in \mathcal{L}(\mathcal{H}), \|U\|_\infty=1} \langle U, \frac{1}{2}\Sigma_\alpha + \frac{1}{2}\Sigma_\beta \rangle \quad (14)$$

$$= \sup_{U=U_1+U_2 \in \mathcal{L}(\mathcal{H}), U_1 \in \text{span}(\Sigma_\alpha), U_2 \in \text{span}(\Sigma_\beta), \|U_1+U_2\|_\infty=1} \frac{1}{2}\langle U_1, \Sigma_\alpha \rangle + \frac{1}{2}\langle U_2, \Sigma_\beta \rangle \quad (15)$$

where by orthogonality we decomposed without loss of generality  $U = U_1 + U_2$  where  $U_1$  and  $U_2$  are restricted to the subspaces defined respectively by  $\Sigma_\alpha$  and  $\Sigma_\beta$  and so mutually orthogonal. Therefore  $\|U_1 + U_2\|_\infty = \max(\|U_1\|_\infty, \|U_2\|_\infty)$  by orthogonality, we can maximise using  $\|U_1\|_\infty = \|U_2\|_\infty = 1$  and recover  $\|\frac{1}{2}\Sigma_\alpha + \frac{1}{2}\Sigma_\beta\|_1 = \frac{1}{2}\|\Sigma_\alpha\|_1 + \frac{1}{2}\|\Sigma_\beta\|_1$ . However, for  $p = 2$ , we get by the definition of the Schatten norm and orthogonality,  $\|\frac{1}{2}\Sigma_\alpha + \frac{1}{2}\Sigma_\beta\|_2^2 = \|\frac{1}{2}\Sigma_\alpha\|_2^2 + \|\frac{1}{2}\Sigma_\beta\|_2^2 = \frac{1}{4}\|\Sigma_\alpha\|_2^2 + \frac{1}{4}\|\Sigma_\beta\|_2^2$ .  $\square$

In particular, when  $\|\Sigma_\alpha\|_2 = \|\Sigma_\beta\|_2$  (for instance,  $\beta$  is a translation of  $\alpha$  and the kernel is translation-invariant),  $\|\frac{1}{2}\Sigma_\alpha + \frac{1}{2}\Sigma_\beta\|_2^2 = \frac{1}{2}\|\Sigma_\alpha\|_2^2$ , there is a decrease by a factor  $\frac{1}{2}$ .

### A.3 Statistical properties (proofs of section 4)

#### A.3.1 Convergence rate (proof of subsection 4.1)

Remember that for clarity of notation, in the proofs we may abbreviate  $\Sigma_\mu$  and  $\Sigma_{\mu_n}$  as  $\Sigma$  and  $\Sigma_n$ .

**Lemma 4.1.** Suppose Assumption 1 and 2 are verified. With a polynomial decay rate of order  $\alpha > 1$  (Assumption P), for  $l = n^{\frac{\theta}{\alpha}}, 0 < \theta \leq \alpha$ :

$$\|P^l(\Sigma_\mu)\Sigma_\mu - \Sigma_\mu\|_1 = R(P^l(\Sigma_\mu)) = \Theta\left(n^{-\theta(1-\frac{1}{\alpha})}\right),$$

$$\|P^l(\Sigma_\mu)\Sigma_\mu - \Sigma_\mu\|_2 = \Theta\left(n^{-\theta(1-\frac{1}{2\alpha})}\right),$$

and there exists  $N \in \mathbb{N}$  such that for  $n > N$ :

$$\|P^l(\Sigma_{\mu_n})\Sigma_\mu - \Sigma_\mu\|_2 \lesssim_{\mu \otimes n} \max(n^{-\frac{1}{2} + \frac{1}{4\alpha}}, n^{-\theta + \frac{1}{4\alpha}}).$$

With an exponential decay rate (Assumption E), for  $l = \frac{1}{\tau} \log n^\theta, \theta > 0$ :

$$\|P^l(\Sigma_\mu)\Sigma_\mu - \Sigma_\mu\|_1 = R(P^l(\Sigma_\mu)) = \Theta(n^{-\theta}),$$

$$\|P^l(\Sigma_\mu)\Sigma_\mu - \Sigma_\mu\|_2 = \Theta(n^{-\theta})$$

and there exists  $N \in \mathbb{N}$  such that for  $n > N$ :

$$\|P^l(\Sigma_{\mu_n})\Sigma_\mu - \Sigma_\mu\|_2 \lesssim_{\mu \otimes n} \begin{cases} \sqrt{\frac{\log n}{n^\theta}} & \text{if } \theta < 1 \\ \frac{(\log n)}{\sqrt{n}} & \text{if } \theta \geq 1. \end{cases}$$

PROOF. The proof of the first point concerning  $R(P^l(\Sigma_\mu))$  for both polynomial and exponential decay can be found for instance in Corollary 3 and 4 of Sterge et al. [2020] which is just bracketing  $R(P^l(\Sigma_\mu)) = \sum_{i>l} \lambda_i = \Theta(\sum_{i>l} f(i))$  by some integrals of  $f$  (which is the function of polynomial or exponential decay).

The proof of the second point concerning  $\|P^l(\Sigma_\mu)\Sigma_\mu - \Sigma_\mu\|_2 = \sqrt{\sum_{i>l} \lambda_i^2}$  is very similar. For the polynomial decay, the  $(\lambda_i)^2$  verify the polynomial decay for  $\alpha' = 2\alpha$ , and  $\theta' = 2\theta$  so that  $0 < \theta' \leq \alpha'$  is equivalent to  $0 < \theta \leq \alpha$  and  $n^{\frac{\theta'}{\alpha'}} = n^{\frac{\theta}{\alpha}} = l$ . Then taking the square root gives the result. Similarly for the exponential decay, the  $(\lambda_i)^2$  verify the exponential decay for  $\tau' = 2\tau$  and  $\theta' = 2\theta$ .

Now for the proof of the third point, denote  $\Sigma_t = \Sigma + tId$  and similarly for  $\Sigma_{n,t} = \Sigma_n + tId$ . Most previous works [Sriperumbudur and Sterge, 2022, Sterge et al., 2020] use what they call  $\mathcal{N}_\Sigma(t) = \text{Tr}(\Sigma(\Sigma + tId)^{-1}) = \|\Sigma^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\|_2^2$ , but for us, we will consider rather  $\|\Sigma\Sigma_t^{-1}\|_2$  and use a result by Rudi et al. [2013]. The proof in the case of the polynomial decay rate concerning

882  $\|P^l(\Sigma_{\mu_n})\Sigma_\mu - \Sigma_\mu\|_2$  or for short  $\|(I - P^l(\Sigma_n))\Sigma\|_2$  goes as the following:

$$\|(I - P^l(\Sigma_n))\Sigma\|_2 = \|(I - P^l(\Sigma_n))\Sigma_{n,t}\Sigma_t^{-1}\Sigma_t\Sigma_t^{-1}\Sigma\|_2 \quad (16)$$

$$\leq \|(I - P^l(\Sigma_n))\Sigma_{n,t}\|_\infty \|\Sigma_t^{-1}\Sigma_t\|_\infty \|\Sigma_t^{-1}\Sigma\|_2. \quad (17)$$

883 We have:

$$\begin{aligned} \|(I - P^l(\Sigma_n))\Sigma_{n,t}\|_\infty &= \hat{\lambda}_{l+1} + t \\ &\leq \frac{3}{2}(\lambda_l + t) \end{aligned} \quad (18)$$

884 with probability  $1 - \delta$  for  $\frac{\kappa}{n} \log(\frac{n}{\delta}) \leq t \leq \lambda_1$  for some constant  $\kappa$  according to Lemma A.1 (iii)  
885 of [Serge et al. \[2020\]](#). Applying Lemma 7.3 of [Rudi et al. \[2013\]](#), we have thanks to the polynomial  
886 decay:

$$\|\Sigma_t^{-1}\Sigma\|_2 = O(t^{-\frac{1}{2\alpha}}). \quad (19)$$

887 Finally, let us show that  $\|\Sigma_{n,t}^{-1}\Sigma_t\|_\infty$  is bounded with high probability for an appropriate range of  $t$ .

888 If we note  $B_n = \Sigma_{n,t}^{-1}(\Sigma_n - \Sigma)$ , then  $\Sigma_{n,t}^{-1}\Sigma_t = Id - B_n$ . Therefore bounding  $\|B_n\|_\infty$  w.h.p. we  
889 can conclude by  $\|\Sigma_{n,t}^{-1}\Sigma_t\|_\infty \leq 1 + \|B_n\|_\infty$ . So to bound  $\|B_n\|_\infty$ , notice that:

$$\|B_n\|_\infty \leq \|\Sigma_{n,t}^{-1}\|_\infty \|\Sigma_n - \Sigma\|_\infty \quad (20)$$

$$= \frac{\|\Sigma_n - \Sigma\|_\infty}{t}. \quad (21)$$

890 Next, we want to apply a concentration bound to  $\|\Sigma_n - \Sigma\|_\infty$  to see what range of  $t$  can be handled.  
891 If we write  $X_k = \frac{1}{n}(\varphi(x_k)\varphi(x_k)^* - \Sigma)$  so that  $\Sigma_n - \Sigma = \sum_{k=1}^n X_k$ , then we have  $\mathbb{E}[X_k] = 0$  and  
892  $\|X_k\|_\infty \leq \frac{2}{n}$ . Besides

$$X_k^2 = \left(\frac{1}{n}\right)^2 (\varphi(x_k)\varphi(x_k)^* - \varphi(x_k)\varphi(x_k)^*\Sigma - \Sigma\varphi(x_k)\varphi(x_k)^* + \Sigma^2)$$

893 since  $\varphi(x_k)\varphi(x_k)^*\varphi(x_k)\varphi(x_k)^* = \varphi(x_k)\varphi(x_k)^*$ , therefore

$$\mathbf{Var}[\Sigma_n - \Sigma] = \sum_{k=1}^n X_k^2 = \frac{1}{n}(\Sigma - \Sigma^2) \preceq \frac{1}{n}\Sigma$$

894 because of the positivity of  $\Sigma^2$ . With all this, we are ready to apply the Matrix Bernstein inequality  
895 for the Hermitian case with intrinsic dimension (Theorem 7.7.1 of [Tropp et al. \[2015\]](#) with  $L = \frac{2}{n}$ ,  
896  $V = \frac{1}{n}\Sigma$ ,  $d = \text{intdim}(V) = \frac{1}{\|\Sigma\|_\infty}$ ,  $v = \frac{\|\Sigma\|_\infty}{n}$ ) and get for  $t' \geq \sqrt{v} + L/3$ :

$$\mathbb{P}(\|\Sigma_n - \Sigma\|_\infty \geq t') \leq \frac{1}{\|\Sigma\|_\infty} \exp\left(\frac{-t'^2/2}{\|\Sigma\|_\infty/n + 2t'/3n}\right) \quad (22)$$

897 So assuming we can pick  $t' = t$  for instance and combining with eq. (21):

$$\mathbb{P}(\|B_n\|_\infty \leq 1) \geq 1 - \frac{1}{\|\Sigma\|_\infty} \exp\left(\frac{-nt^2/2}{\|\Sigma\|_\infty + 2t/3}\right) \quad (23)$$

898 Let us pick  $t = \frac{K}{\sqrt{n}}$  for some  $K$  big enough, so that the condition on  $t'$  and the condition for eq. (18)  
899 to work are satisfied and the exponential term in eq. (23) is as small as desired to make the bound  
900 sensible (it is true for  $n$  big enough). Combining the latter eq. (23) with eq. (18) and (19), eq. (17)  
901 gives

$$\|(I - P^l(\Sigma_n))\Sigma\|_2 \lesssim_{\mu^{\otimes n}} t^{-\frac{1}{2\alpha}}(\lambda_{l+1} + t)$$

902 and replacing  $t = \frac{K}{\sqrt{n}}$  and  $\lambda_{l+1} = \Theta((l+1)^{-\alpha}) = \Theta(n^{-\theta})$ :

$$\|(I - P^l(\Sigma_n))\Sigma\|_2 \lesssim_{\mu^{\otimes n}} n^{-\frac{1}{2} + \frac{1}{4\alpha}} + n^{-\theta + \frac{1}{4\alpha}}$$

903 hence the result.

904 Now for the case of exponential decay rate, since it is a more powerful hypothesis we use a simpler  
905 argument:

$$\begin{aligned} \|(I - P^l(\Sigma_n))\Sigma\|_2 &\leq \|(I - P^l(\Sigma_n))\Sigma^{1/2}\|_2 \|\Sigma^{1/2}\|_\infty \\ &\leq \|(I - P^l(\Sigma_n))\Sigma^{1/2}\|_2 \end{aligned}$$

906 since  $\|\Sigma\|_1 = 1$ , and it turns out that  $\|(I - P^l(\Sigma_n))\Sigma^{1/2}\|_2 = \sqrt{R(P^l(\Sigma_n))}$ . According to Sterge  
 907 et al. [2020],

$$R(P^l(\Sigma_n)) \lesssim_{\mu^{\otimes n}} \begin{cases} \frac{\log n}{n^\theta} & \text{if } \theta < 1 \\ \frac{(\log n)^2}{n} & \text{if } \theta \geq 1 \end{cases} \quad (24)$$

908 hence the result.  $\square$

909 By the work of Blanchard et al. [2007], we can state one of their results in a simplified lemma:

910 **Lemma A.3.** [Blanchard et al. 2007] Suppose Assumption 1 and 2 are verified. For all projector  $P$   
 911 of rank  $l$ , with probability at least  $1 - e^{-\xi}$ :

$$R_n(P) \leq \frac{3}{2}R(P) + 24\sqrt{\frac{l}{n}(1 - \|\Sigma_\mu\|_2^2)} + \frac{25\xi}{n} \quad (25)$$

912 This comes from their equation (30), with  $M = 1$  in our case, using  $K = 2$  and bounding their  
 913  $\rho(M, l, n)$  by  $\sqrt{\frac{l}{n} \text{tr} C'_2}$  they mentioned as they suggested (where they described  $C'_2 = \int_{\mathcal{X}} (\varphi(x) \otimes$   
 914  $\varphi(x)^*)^* \otimes (\varphi(x) \otimes \varphi(x)^*) d\mu(x) - \Sigma_\mu \otimes \Sigma_\mu$ ). For us, the only important point is that it is a bounded  
 915 quantity.

916 **Theorem 4.2.** Suppose Assumption 1 and 2 are verified.

- 917 • If the eigenvalues of  $\Sigma_\mu$  follow a polynomial decay rate of order  $\alpha > 1$  (Assumption P),  
 918 then:

$$d_{KT}(\mu, \mu_n) \lesssim_{\mu^{\otimes n}} n^{-\frac{1}{2} + \frac{1}{2\alpha}}.$$

- 919 • If the eigenvalues of  $\Sigma_\mu$  follow an exponential decay rate (Assumption E), then:

$$d_{KT}(\mu, \mu_n) \lesssim_{\mu^{\otimes n}} \frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}.$$

920 **PROOF.** By the triangular inequality:

$$\begin{aligned} \|\Sigma - \Sigma_n\|_1 &\leq \|\Sigma - P^l(\Sigma)\Sigma\|_1 + \|(P^l(\Sigma) - P^l(\Sigma_n))\Sigma\|_1 + \|P^l(\Sigma_n)(\Sigma - \Sigma_n)\|_1 + \|P^l(\Sigma_n)\Sigma_n - \Sigma_n\|_1 \\ &\leq R(P^l(\Sigma)) + \sqrt{2l}\|(P^l(\Sigma) - P^l(\Sigma_n))\Sigma\|_2 + \sqrt{l}\|(\Sigma - \Sigma_n)\|_2 + R_n(P^l(\Sigma_n)) \\ &\leq R(P^l(\Sigma)) + \sqrt{2l}(\|(P^l(\Sigma)\Sigma - \Sigma\|_2 + \|P^l(\Sigma_n)\Sigma - \Sigma\|_2) + \sqrt{l}\|(\Sigma - \Sigma_n)\|_2 + R_n(P^l(\Sigma_n)) \end{aligned} \quad (26)$$

921 where we have mainly used Schatten norm ‘‘H lder’’ inequality between 1-norm and 2-norms (eq. (5))  
 922 as well as  $\|\Sigma - P^l(\Sigma)\Sigma\|_1 = \sum_{i>l} \lambda_i = R(P^l(\Sigma))$  and similarly for  $\Sigma_n$ . Now, thanks to Lemma 4.1,  
 923 for the polynomial case:

924 • 
$$R(P^l(\Sigma)) = \Theta\left(n^{-\theta(1 - \frac{1}{\alpha})}\right).$$

- 925 • By Lemma A.3

$$\begin{aligned} R_n(P^l(\Sigma_n)) &\leq R_n(P^l(\Sigma)) \\ &\lesssim_{\mu^{\otimes n}} R(P^l(\Sigma)) + \sqrt{\frac{l}{n}} + \frac{1}{n} \\ &\lesssim_{\mu^{\otimes n}} \max(R(P^l(\Sigma)), \sqrt{\frac{l}{n}}) \\ &\lesssim_{\mu^{\otimes n}} \max(n^{-\theta(1 - \frac{1}{\alpha})}, n^{-\frac{1}{2} + \frac{\theta}{2\alpha}}). \end{aligned}$$

926 • 
$$\begin{aligned} \sqrt{2l}\|(P^l(\Sigma)\Sigma - \Sigma\|_2 &\lesssim_{\mu^{\otimes n}} n^{\frac{\theta}{2\alpha}} n^{-\theta + \frac{\theta}{2\alpha}} \\ &= n^{-\theta + \frac{\theta}{\alpha}}. \end{aligned}$$

927 • 
$$\begin{aligned} \sqrt{2l}\|(P^l(\Sigma_n)\Sigma - \Sigma\|_2 &\lesssim_{\mu^{\otimes n}} n^{\frac{\theta}{2\alpha}} \max(n^{-\frac{1}{2} + \frac{1}{4\alpha}}, n^{-\theta + \frac{1}{4\alpha}}) \\ &= \max(n^{-\frac{1}{2} + \frac{1}{4\alpha} + \frac{\theta}{2\alpha}}, n^{-\theta + \frac{\theta}{2\alpha} + \frac{1}{4\alpha}}). \end{aligned}$$

928

- From MMD convergence rates, we know:

$$\begin{aligned}\sqrt{l}||(\Sigma - \Sigma_n)||_2 &\lesssim_{\mu^{\otimes n}} \sqrt{\frac{l}{n}}. \\ &= n^{-\frac{1}{2} + \frac{\theta}{2\alpha}}\end{aligned}$$

929 Now combining all those terms, from eq. (26) we get, for  $0 < \theta \leq \alpha$ :

$$\begin{aligned}||\Sigma - \Sigma_n||_1 &\lesssim_{\mu^{\otimes n}} n^{-\max(-\theta + \frac{\theta}{\alpha}, -\frac{1}{2} + \frac{\theta}{2\alpha}, -\frac{1}{2} + \frac{1}{4\alpha} + \frac{\theta}{2\alpha}, -\theta + \frac{\theta}{2\alpha} + \frac{1}{4\alpha})} \\ &\lesssim_{\mu^{\otimes n}} n^{-\max(-\theta + \frac{\theta}{\alpha}, -\frac{1}{2} + \frac{1}{4\alpha} + \frac{\theta}{2\alpha}, -\theta + \frac{\theta}{2\alpha} + \frac{1}{4\alpha})}.\end{aligned}$$

930 The decreasing lines  $-\theta + \frac{\theta}{\alpha}$  and  $-\theta + \frac{\theta}{2\alpha} + \frac{1}{4\alpha}$  and the increasing line  $-\frac{1}{2} + \frac{1}{4\alpha} + \frac{\theta}{2\alpha}$  all cross at  
931  $\theta = \frac{1}{2}$ , so we find it is the optimal  $\theta$  to minimise, which gives a rate of  $n^{-\frac{1}{2} + \frac{1}{2\alpha}}$ .

932 Similarly, we do the same for the exponential decay rate case, in particular if we take  $\theta = 1$  we get:

$$||\Sigma - \Sigma_n||_1 \lesssim_{\mu^{\otimes n}} \frac{(\log n)^{\frac{3}{2}}}{\sqrt{n}}$$

933 where the dominating term in eq. (26) is  $\sqrt{2l}||(\Sigma - \Sigma_n)||_2$ . □

### 934 A.3.2 Robustness properties (proof of subsection 3.3)

935 **Proposition 3.6** Denote  $P_\varepsilon = (1 - \varepsilon)P + \varepsilon C$  where  $C$  is some contamination distribution. We  
936 have when Assumption 1 is verified:  $|d_{KT}(P_\varepsilon, Q) - d_{KT}(P, Q)| \leq 2\varepsilon$ .

937 **PROOF.** Since  $d_{KT}$  is a metric based on a norm:

$$\begin{aligned}|d_{KT}(P_\varepsilon, Q) - d_{KT}(P, Q)| &\leq d_{KT}(P, P_\varepsilon) \\ &= ||\varepsilon(\Sigma_P - \Sigma_C)||_1 \\ &= \varepsilon||(\Sigma_P - \Sigma_C)||_1 \\ &\leq 2\varepsilon\end{aligned}$$

938 □

939 The proof works for the Schatten 2-norm ( $MMD_{k^2}$ ) as well.

940 On the contrary in  $(\mathbb{R}^d, || \cdot ||)$  for the classical Wasserstein 1-distance, if for distribution  $Q$ , we have  
941  $1 - \frac{\varepsilon}{2}$  mass contained in a ball of some center  $c$  and some radius  $r$ , then taking as contamination  
942  $C = \delta_x$  a Dirac in some point  $x$ , we must have  $W_1(P_\varepsilon, Q) \geq \frac{\varepsilon}{2}||x - c|| - r$  and we can take  
943  $||x|| \rightarrow \infty$  to make the distance diverge.

### 944 A.4 Proof of Proposition 2.3

945 **PROOF.** Notice that, as  $\tilde{\varphi}(z_k)\tilde{\varphi}(z_k)^* = (\mu_n - \nu_m)(z_k)\varphi(z_k)\varphi(z_k)^*$ , we have  $\Sigma_{\mu_n - \nu_m} =$   
946  $\sum_{k=1}^r \tilde{\varphi}(z_k)\tilde{\varphi}(z_k)^* = ZZ^*$ . Since  $\Sigma_{\mu_n - \nu_m}$  is real symmetric, it is diagonalisable, and as a sum of  $r$   
947 projectors, it is of rank  $r$  (by linear independence of the  $(\varphi(z_k))_{i=1\dots r}$ ). Let us denote  $(\lambda_k, v_k)_{k=1, \dots, r}$   
948 the couples of eigenvalues and eigenvectors of the restriction of  $\Sigma_{\mu_n - \nu_m}$  on the subspace of dimension  
949  $r$  spanned by those projectors. Those eigenvectors are orthogonal and those eigenvalues are non-zero.  
950 Note that for such an eigen-(value, vector)  $(\lambda, v)$  of  $ZZ^*$ , we have:  $Z^*ZZ^*v = Z^*(\lambda v)$ , therefore  
951  $Z^*v$  is an eigenvector of  $Z^*Z$  with associated eigenvalue  $\lambda$  as well. Then note that  $(Z^*v_k)_{k=1, \dots, r}$  are  
952 also orthogonal and their norm is not zero, since  $\langle Z^*v_i, Z^*v_j \rangle = v_i^*ZZ^*v_j = v_i^*\lambda_j v_j = \lambda_j \langle v_i, v_j \rangle$ .  
953 Therefore they are distinct, and we can fully diagonalise  $K$  (which is of size  $r \times r$ ) with such  
954 vectors. □

## 955 B EXPERIMENTS

956 Computation were carried on a Macbook Pro 2020 with processor 2,3 GHz Intel Core i7 (4 cores)  
957 and memory 16 Go 3733 MHz LPDDR4X (graphic card is Intel Iris Plus Graphics 1536 Mo).

## B.1 Normalised energy (Additional figures for subsection 3.2)

Here are some other simulations to highlight the different geometrical behaviours of MMD and  $d_{KT}$ : in Figure 2, we compute the two distances (both with Gaussian kernel with  $\sigma = 1$ ) for two sets of  $n = 1000$  samples, one following a standard normal law  $\mathcal{N}(0, 1)$  and the other following  $\mathcal{N}(\theta, 1)$  and we try for different values of  $\theta$  from 0 to 10 with steps of 0.5. As can be seen, even when the two locations are far apart, since the two distributions are not Dirac, their variances prevent MMD to reach the maximum 2. This means that when the distance is getting close to zero in general the slope will be flatter for MMD than for  $d_{KT}$ : one should be aware of that when doing for instance a gradient descent. We also added the Kernel Bures-Wasserstein (computed with a simplified kernel trick as in Oh et al. [2020]) to illustrate the Fuchs-van de Graaf inequalities. For  $\theta$  high enough, the lower bound joins  $d_{KT}$  (they naturally cannot go higher than 2 because of Assumption 1).

In Figure 3 this time, the same experiment is shown but this time between distributions  $\mathcal{N}(0, s)$  and  $\mathcal{N}(100, s)$ , where we make the variance  $s$  varies in  $[0.1, 0.3, 1, 3, 10, 30, 100]$ . As the variance grows, the Hilbert norm of the distributions decreases and so the MMD decreases very fast. On the contrary,  $d_{KT}$  takes its maximum value 2 distinguishing the far-away distributions and starts only to decrease for high values of  $s$  down to close to 1 for  $s = 100$  (which the distance between the two locations), which seems reasonable as, for low variance values, the support of the majority of the central masses of each distributions do not even overlap!

## B.2 ABC (Additional table and figures)

In Table 2, we display more complete results of the experiments adding other methods as well as the standard deviations over the different runs. Referring to the distances of Corollary 3.4, we added the Optimal Transport (OT) 1-Wasserstein distance with Euclidean cost (as we mentioned it rejects all samples because of the contamination), as well as the one with a distance cost based on the Gaussian kernel  $c_k$  (noted as  $OT_{gauss}$  in the table) and the Kernel Bures-Wasserstein distance ( $d_{KBW}$ ). We added some “normalised” MMD:

$$MMD_N(\mu, \nu) = \frac{\|\Sigma_\mu - \Sigma_\nu\|_2}{\sqrt{\|\Sigma_\mu\|_2^2 + \|\Sigma_\nu\|_2^2}}$$

from the inequality of Section 3.2 in order to attenuate its effect. It is not guaranteed to be a distance, and eventually it does not perform better than  $d_{KT}$ . For the sake of fairness, we compare ourselves to competitors that also require cubic complexity of computation such as the Kernel Fischer Discriminant Analysis (KFDA [Eric et al. 2007a]) with parameter  $\gamma_n = n^{-1/2} = 0.1$  but it also rejects all the time, so we added its normalised version as well.

In Figure 4 are displayed the simulated posteriors  $\frac{1}{|L_\theta|} \sum_{\theta_i \in L_\theta} p(\cdot | \theta_i)$  obtained as a result of Rejection ABC Algorithm 1 using the Gaussian kernel for both MMD and  $d_{KT}$ . For sensible parameter  $\varepsilon$ , the posteriors obtained via  $d_{KT}$  are quite close to the target, while for for MMD (gaussian) the posterior stays very flat like the prior unless  $\varepsilon$  goes to a very low value (less than the contamination threshold). For sake of visibility, the best posteriors of normalised MMD and MMD with the energy kernel are displayed in another Figure 5. For the energy kernel, we can see that the peak of the density is not aligned with the one of the target.

---

### Algorithm 1 Rejection ABC Algorithm

---

**Require:** Observed data  $\{X_i\}_{i=1}^n$ , prior  $\pi(\theta)$  on the parameter space  $\Theta$ , tolerance threshold  $\epsilon$ , statistical distance  $d$ , empty list  $L_\theta$

```

1: for  $i = 1$  to  $T$  do
2:   draw  $\theta_i \sim \pi(\theta)$ 
3:   draw  $Y_1 \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} p_{\theta_i}$ 
4:   if  $d(X^n, Y^m) < \epsilon$  then
5:     Add  $\theta_i$  to  $L_\theta$ 
6:   end if
7: end for
8: return  $L_\theta$ 
```

---

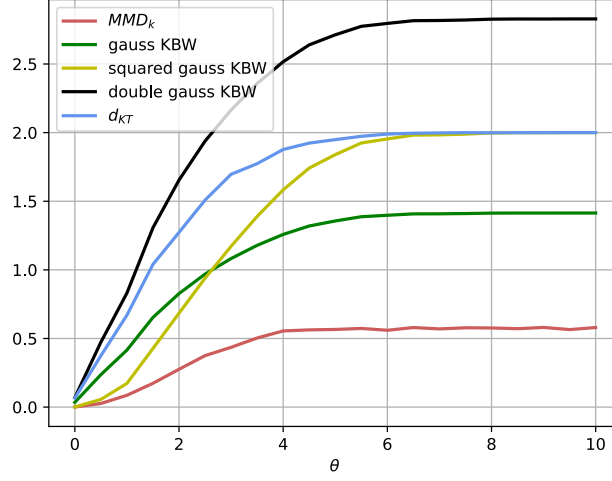


Figure 2: Variations on the mean  $\theta$

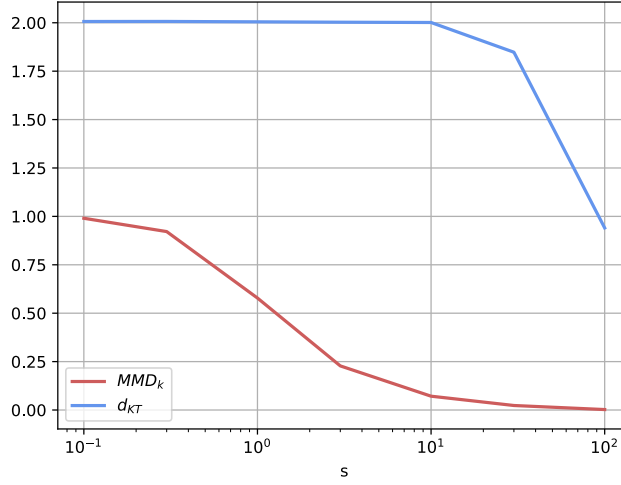


Figure 3: Variations on the standard deviation  $s$

### B.3 Particle gradient flows

Here are displayed at different iterations the particle flows of  $d_{KT}$  (Fig. 6) and MMD (Fig. 7). We choose the Laplacian kernel over the Gaussian kernel, as it gave a better convergence results for both  $d_{KT}$  and MMD. Even though it is not differentiable at the coordinates of the target particles, in practice computationally we observed that no problem occurred as this typically would happen when the cloud of points are reaching destination.

We also add a shape transfer task as in Chazal et al. [2024] displayed in Figure 8 where we added to their results our  $d_{KT}$  using the Laplacian kernel with bandwidth  $\sigma = 1$  again and with a learning rate of 5. The other methods comprises the  $KKL_\alpha$  with  $\alpha = 0.01$  which is a regularised version of the KKL which uses the same RKHS density operators as us but using the Von Neumann quantum relative entropy and is also in cubic time complexity, and KALE with parameter  $\lambda = 0.001$  which approaches the classic KL divergence, and  $\lambda = 10000$  which approaches MMD (and thus we call it so in the figure).



Table 2: Average MSE of ABC Results.

The Optimal Transport Wasserstein distance (OT) has been added, which rejects all samples. The KFDA is used with  $\gamma_n = n^{-1/2} = 0.1$ .

$\varepsilon$	distance	#accept. (std)	MSE (std)
0.05	OT	0	N/A
	KFDA	0	N/A
	KFDA <sub>norm.</sub>	1457 (123)	0.41 (0.05)
	MMD	1092 (45)	0.19 (0.02)
	MMD <sub>N</sub>	0	N/A
	MMD <sub>E</sub>	0	N/A
	$d_{KBW}$	0	N/A
	OT <sub>gauss</sub>	0	N/A
	$d_{KT}$	0	N/A
0.25	OT	0	N/A
	KFDA	0	N/A
	KFDA <sub>norm.</sub>	1557 (122)	0.45 (0.05)
	MMD	2964 (92)	1.29 (0.06)
	MMD <sub>N</sub>	840 (30)	0.12 (0.01)
	MMD <sub>E</sub>	0	N/A
	$d_{KBW}$	0	N/A
	OT <sub>gauss</sub>	343 (48)	0.04 (0.01)
	$d_{KT}$	58 (25)	<b>0.03</b> (0.01)
0.5	OT	0	N/A
	KFDA	0	N/A
	KFDA <sub>norm.</sub>	1673 (118)	0.49 (0.05)
	MMD	6168 (406)	7.47 (1.83)
	MMD <sub>N</sub>	1964 (69)	0.57 (0.02)
	MMD <sub>E</sub>	846 (35)	0.17 (0.05)
	$d_{KBW}$	1312 (49)	0.26 (0.02)
	OT <sub>gauss</sub>	1376 (53)	0.29 (0.02)
	$d_{KT}$	828 (34)	0.12 (0.01)
1	OT	0	N/A
	KFDA	0	N/A
	KFDA <sub>norm.</sub>	1847 (121)	0.57 (0.05)
	MMD	10000 (0)	26.0 (0.18)
	MMD <sub>N</sub>	9488 (57)	20.4 (0.31)
	MMD <sub>E</sub>	2926 (52)	1.33 (0.6)
	$d_{KBW}$	3709 (54)	2.02 (0.05)
	OT <sub>gauss</sub>	3484 (84)	1.78 (0.06)
	$d_{KT}$	2067 (93)	0.63 (0.04)

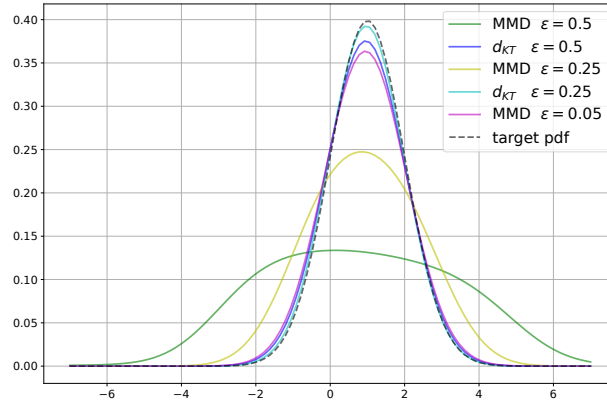


Figure 4: Posterior probability density functions using Gaussian kernel

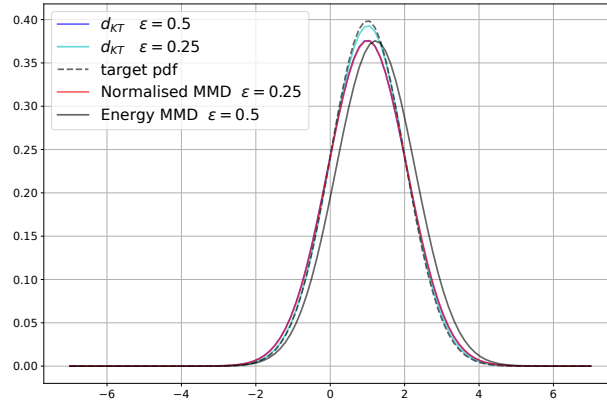


Figure 5: Posterior probability density functions of  $d_{KT}$  and other competitors than classic MMD Gaussian kernel

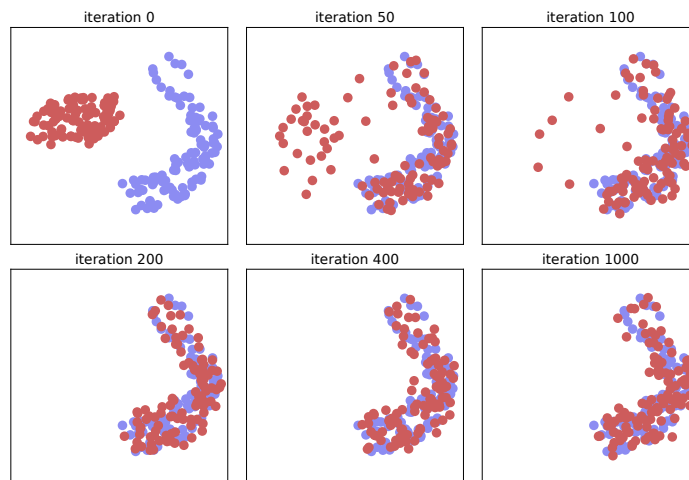


Figure 6: Particle flow with  $d_{KT}$  leads to a good match between the distributions

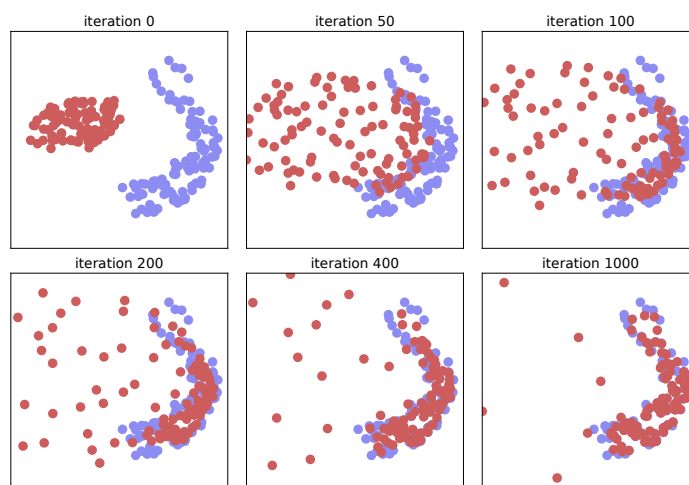


Figure 7: Particle flow with MMD leads to several samples being “repulsed” due to internal energy.

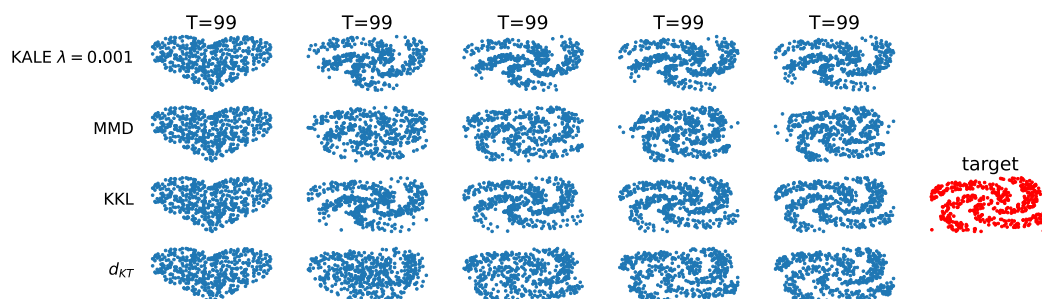


Figure 8: Shape transfer