# A Fast, Consistent Kernel Two-Sample Test: Appendix

**Arthur Gretton**
Carnegie Mellon University
MPI for Biological Cybernetics
*arthur.gretton@gmail.com*

**Kenji Fukumizu**
Inst. of Statistical Mathematics
Tokyo Japan
*fukumizu@ism.ac.jp*

**Zaid Harchaoui**
Carnegie Mellon University
Pittsburgh, PA, USA
*zaid.harchaoui@gmail.com*

**Bharath K. Sriperumbudur**
Dept. of ECE, UCSD
La Jolla, CA 92037
*bharathsv@ucsd.edu*

## 1 Proof of convergence of covariance operator trace

We begin with a number of standard definitions and results, taken from [1]. We recall first the definition of the covariance operator for a distribution $P$ on $\mathcal{X}$ from equation (4) in the main document. This can be written

$$C := \widetilde{C} - M,$$

where $\tilde{C} := \mathbf{E}(\phi(x) \otimes \phi(x))$ is the uncentered covariance operator, $M := \mu \otimes \mu$, and we have defined $f \otimes g : \mathcal{H} \to \mathcal{H}$ such that $(f \otimes g)h = \langle g, h \rangle_{\mathcal{H}} f$. Next, given any orthonormal basis $\psi_i$ for $\mathcal{H}$, the trace of $C$ is

$$\mathrm{tr}(C) := \sum_{p=1}^{\infty} \langle \psi_p, C\psi_p \rangle .$$

In particular,

$$\mathrm{tr}(f \otimes g) = \langle f, g \rangle ,$$

from which it follows that $\mathrm{tr}(C) := E\left\|\phi(x)\right\|^2 - \left\|\mu\right\|^2$.

We next describe the empirical covariance operator based on a sample $(x_1, \ldots, x_m)$, and specify its trace. We write

$$C_m := \widetilde{C}_m - M_m,$$

where

$$\widetilde{C}_m := \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \otimes \phi(x_i)$$

and

$$M_m := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \phi(x_i) \otimes \phi(x_j).$$

It follows that

$$\mathrm{tr}(C_m) = \frac{1}{m} \sum_{i=1}^{m} \left\|\phi(x_i)\right\|^2 - \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \langle \phi(x_i), \phi(x_j) \rangle .$$

Before proceeding to our main result, we recall McDiarmid's theorem [2].

**Theorem 1 (McDiarmid)** *Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, m\}$, there exist $c_i < \infty$ for which*

$$\sup_{\mathbf{x} \in \mathcal{X}^m, \tilde{x} \in \mathcal{X}} |f(x_1, \ldots, x_m) - f(x_1, \ldots x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)| \leq c_i.$$

*Then for all measures $P$ on $\mathcal{X}$ and every $t > 0$,*

$$P_{x^m}\left(f(\mathbf{x}) - \mathbf{E}_{x^m}(f(\mathbf{x})) > t\right) < \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right).$$

We now proceed to our main result.

**Theorem 2** *Assume $\|\phi(x)\|^2 \leq B$ for all $x \in \mathcal{X}$. Then $|\mathrm{tr}(C) - \mathrm{tr}(C_m)| = o_p(m^{-1/2})$.*

**Proof** First, we decompose

$$|\mathrm{tr}(C) - \mathrm{tr}(C_m)| \leq \left|\mathrm{tr}(\widetilde{C}) - \mathrm{tr}(\widetilde{C}_m)\right| + |\mathrm{tr}(M) - \mathrm{tr}(M_m)|.$$

We consider the first term. Define as $\widetilde{C}'_m$ the empirical covariance operator obtained by replacing the $i_0$th sample $x_{i_0}$ with $x'_{i_0}$. Then

$$\left|\mathrm{tr}\widetilde{C}_m - \mathrm{tr}\widetilde{C}'_m\right| = \frac{1}{m}\left|\|\phi(x_{i_0})\|^2 - \left\|\phi(x'_{i_0})\right\|^2\right| \leq \frac{B}{m},$$

and we obtain convergence in probability with rate $m^{-1/2}$ by McDiarmid's theorem. We next consider the second term. Again replacing a particular $x_{i_0}$ with $x'_{i_0}$ and defining the resulting empirical centering operator as $M'_m$, we get[1]

$$\begin{aligned} |\mathrm{tr}M_m - \mathrm{tr}M'_m| &\leq \frac{2}{m(m-1)}\left|\sum_{j \neq i_0}\left\langle \phi(x_{i_0}) - \phi(x'_{i_0}), \phi(x_j)\right\rangle\right| \\ &\leq \frac{2}{m(m-1)}\sum_{j \neq i_0}\left\|\phi(x_{i_0}) - \phi(x'_{i_0})\right\| \|\phi(x_j)\| \\ &\leq \frac{4B}{m}. \end{aligned}$$

We apply McDiarmid's theorem to obtain convergence in probability with rate $m^{-1/2}$.

■

## References


[1] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294, 2007.

[2] C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.


[1]Note: the sums $\sum_{j \neq i_0}$ in the expressions below are taken only over the index $j$, and not the fixed index $i_0$.

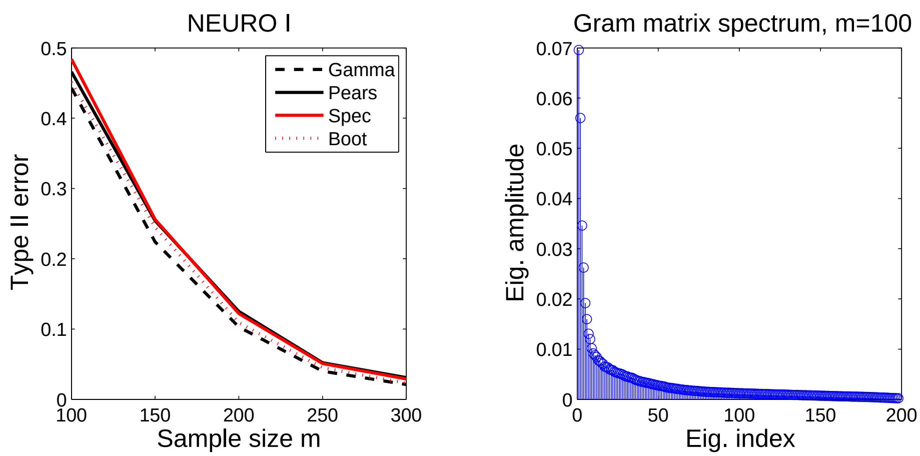


Figure 1: NEURO I dataset. **Left:** Plot of Type II error vs number $m$ of samples. **Right:** Eigenspectrum of a centered Gram matrix obtained by drawing $m = 100$ points from each of $P$ and $Q$, where $P \neq Q$.