**Comments that I don’t know what to do with**

So far mainly asymptotic theory is given; it is a pity that there is no discussion of the limits.

A real practical application is not included (so it remains unclear whether the assumption of the theorems would be fulfilled in real data)

also a conclusion is missing.

It would be nice if the exposition and notation could be improved for the paper to be more easily accessible by a larger part of the community.

Use of the word "control": please correct me if I am wrong, but in my understanding the type 1 error is controlled if the truly measured type 1 error is \*smaller or equal\* to the desired level. In several experiments (e.g., MCMC Vb1) the actual rate is slighly higher than the desired one, which would strikly speaking mean the error is not entirely controlled (if I get the terminology right).

- For a sound empirical analysis, you could use ROC curves. It is neccessary to compare the type 1 error and the power / type 2 error at various level simultaeoulsy for a fair comparision.

2. What does Figure 1 tell me? The methods could be better evaluated if you plotted a roc or pr curve or something like that, I would think.

Same page: "The type... was equal or lower ..." This is again vague. A roc or pr curve would shed light on this.

- The Bonferoni correction for HSIC-Lag will be too conservative under the alternative hypothesis. Is there a way to wrap a (wild) bootstrap around the whole lagged kind of procedure / come up with a joint statistics, which is calibrated via wild bootstrap?

- experiments: make clearer what the difference between MMD hat and Vb1 and Vb2 is. Also note that the Type 1 error is not controlled, it is just closer to the desired error

**Comments that I can handle.**

**yellow not done**, due to lack of space

expectation could take the usual symbol instead of the current one

filtration could be explained/defined in more details (section 2) [ignored due to lack of space]

MMD\_k in proposition 1 is not defined (it is, though, in proposition 5 in the appendix)...

For instance, for \tau-mixing very little explanation is provided (also some consequences of these assumptions would be appreciated), bootstrap assumptions are mentioned only very briefly and raise many questions for the reader.

In the last paragraph of the Section 2, authors also use `canonical cores of degree m=2' without explicitly defining what the degree of a core is [ignored - such a lie ! ]

In the definition of the V-statistic there is a Z\_0 variable which is never introduced

Authors could also provide a sentence or two on Prokhorov metric. [ignored - lack of space]

The authors are asked to briefly discuss the computational complexity compared to alternative approaches. It seems that it does not have any additional cost when compared to other possible alternatives such as the permutation test described in [7].

It is recommended to augment the paper with a tailored introduction to the wild bootstrap - perhaps as supplementary material?

- Introduction of the wild bootstrap at the end of Sect 2: this is very technical, but nobody in ML knows this quantity. Show an illustration, what does it do intuitively? Example of how to choose the wild bootstrap process: for instance, the statistical learning theory reader might wonder whether it makes sense to use a correlated Rademacher or Gaussian process here.

- Beginning of Sec 4: "This quantity". Let's say I wouldn't know anything about MMD: it is unclear to which term "this quantity" refers.

- Top of page 5: "[18] and 2" -> [18] and [25] ?

1. at some places in the text (e.g., end of page 7) you mention the measured type 2 error. This is important and I would suggest mentioning this also in the captions of the figures.

3. For the same reason: figure 2 and end of page 8. "was around 5%". This is a little vague and again this should be mentioned in the caption of figure 2.