We thank the reviewers for their thorough reading of our work.

We thanks Rev 1 for pointing out the results of the preprint [3]. Indeed [3] study a class of two sample test statistics based on inter-point distances and they show benefits of using the ℓ_1 norm over the Euclidean distance and the Maximum Mean Discrepancy (MMD) when the dimensionality goes to infinity. For this class of test statistics, they characterize asymptotic power loss w.r.t the dimension and show that the ℓ_1 norm is beneficial compared to the ℓ_2 norm and the MMD provided that the summation of discrepancies between marginal univariate distributions is large enough. These results echo our message that ℓ_1-based tests outperform their ℓ_2 counterparts and we will mention them in the final version. Our work is complementary and differs on several aspects. The population version of the ℓ_1-based statistic ($\mathbf{ED}^1$) proposed in [3] does not fully characterize the difference between two distributions P, Q since $\mathbf{ED}^1(P, Q) = 0$ does not necessarily imply $P = Q$ while the population version of our ℓ_1-based tests are metric on the space of Borel probability measures which metrize the weak convergence. Our experiment varying the dimensionality of the problem is the part of our work that is closest to [3], and confirms their general message, with a different test procedure. However, the test statistics introduced in [3] have a quadratic cost in the sample size and the tests realized are permutation based tests, which can be very expensive when the sample size increases while our tests are linear in the sample size. With regards to the link to [1], our work adapts the idea of deriving (Monte Carlo) approximation to our ℓ_1 metric, while [1] had studied a related approximation in the Euclidean case. It also establishes many more results on the ℓ_1-based metric, including the IPM formulation, the weak convergence properties, and the lower bound on the test power. Moreover we show both theoretically and empirically the benefits of using the ℓ_1 norm compared to the Euclidean geometry.

Following the suggestion of the Rev 2, we will explain the exact parameter settings, which is the same as in [2], that we used for the initialization of the test locations, the Gaussian width and the value used for regularization parameter γ_{N_1,N_2} to compute the optimized tests **L1-opt-ME** and **L1-opt-SCF**. The statistics presented in our paper capture differences between distribution representatives in a RKHS at J locations. When using analytic kernels, these differences between representatives become dense. Therefore when the sample size is large enough, these dense differences are large enough to allow the ℓ_1-norm to reject better the null hypothesis with high probability. The most important weakness of our study is probably the optimization of the cost functions. The lower bound that we optimize is non-convex, as in the ℓ_2 prior art [2]. However, the use of the ℓ_1-norm makes optimization even harder, as it is no longer a smooth. Further work should consider dedicated optimization algorithms. We will try to add a mention of this weakness in the manuscript. In Table 1, we run the different optimized tests on the Blobs problem when the test sample size is $n^{te} = 1e6$.

Table 1: Higgs dataset: Table of the running times of the optimized tests when $n^{te} = 1e6$ and $J = 2$.

	L1-opt-ME	**ME-full**	**L1-opt-SCF**	**SCF-full**
Running Time (s)	164.23	157.97	599.77	579.42

To answer Rev 3, in practice, the distribution Γ that we use to sample the $\{T_j\}_{j=1}^J$ to compute the grid version of our statistics, **L1-grid-ME** and **L1-grid-SCF**, is a multivariate normal distribution (see line 197). More specifically, for both tests, we sample the test locations with realizations from two multivariate normal distributions fitted to samples from P and Q; this ensures that the locations are well supported by the data. By denoting $t = N_1 + N_2$, for both tests the testing cost is $\mathcal{O}(J^3 + Jt + dJt)$ and the optimization costs is $\mathcal{O}(J^3 + dJt)$ per gradient ascent (see line 214 - 215). In the second panel of Figures 1 and 2, the ME lines are drawn but their reach the minimum Type-II error of 0.0 when the dimension is between 5 and 1500 or the sample test size is between 500 and 5500. Indeed the GMD problem is an easy problem for this test. Therefore we decide to plot in the supp. mat. an harder version of this problem where the difference between means is smaller (see Figure 4). In Table 2, the McDo vs McDo problem corresponds to the Type-I errors and the other problems represent Type-II errors. We will follow the suggestion proposed by Rev 3 and create two separate tables for the Type-I errors and Type-II error respectively. We will also add the pdf of the Nakagami distribution in the main text. To show experimentally the results of the Proposition 3.1 and 3.4, we need to compute the quantiles of the asymptotic null distributions of the different unnormalized tests that we have presented which require a computationally-costly bootstrap or permutation procedure. As we do not have computational ressources available before the end of the author response period to run a full experiment, we prefer to provide such results later but they will be added in the final manuscript. We will also try to release the code as early as possible.

References

[1] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, page 1981, 2015.

[2] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

[3] C. Zhu and X. Shao. Interpoint distance based two sample tests in high dimension. *arXiv:1902.07279*, 2019.