Review Report

Thanks to the revision and the reply letter, I think now I have a better understanding of this manuscript. This paper can be divided into two parts: The first part introduces a unifying framework of upper-bounds (called minimum risks) on the Bayes error with respect to various loss functions $\phi$. Arbitrary tighter bounds are also derived using polynomial functions. The second part illustrates the connection between the minimum risk and the concept of refinement. This idea leads to the development of a projected discrepancy between two distributions in an embedded space.

The technical content of this paper looks sound and this version has a better readability than the previous submission, however, some of the reasoning are still hard to understand thus I recommend authors keep improving the presentation of this paper. Here I list a few specific concerns:

1) The motivation of using minimum risk minimization for hypothesis test is not clear. In this paper, the idea of “minimum risk” is introduced as an upper-bound of “Bayes error” and it makes me feel this is an ideal tool for solving classification problems. However, authors quickly turned to the discussion of two sample test problems without a strong motivation (Section 4). In abstract, author mentioned “…showing that the minimum risk is closely related to the concept of refinement. This directly leads to the definition of the minimum projected minimum risk discrepancy and the kernelized minimum projected minimum risk discrepancy as a means of measuring distances between embedded distributions in a kernel space...”. Frankly, after reading the paper, I am still confused that how the relation to refinement leads to a new discrepancy measure and why the minimum risk is a better statistic for two sample test problems than the other possible measures. Authors should emphasis their motivations in the earlier sections of this paper.

   i. In page 15, authors mentioned finding $w^*$ is difficult and then focusing on special cases have closed forms and convex solutions and we can obtain a projection $w$ by solving SVM or Fisher Discriminant Analysis. However, do they minimize the minimum risk in some sense? If not how to justify the usage of these $w$s? If there are many other $w$ that preserve the injective property, how do we choose?

2) After deriving an arbitrary upper-bound of Bayes error, authors suggest itself can be used for many applications (page 11). However, to compute such a bound (as described in Table 4), we need to the access to $P(x|1)$ and $P(x)$. If we have already known the generating sources (generative models), why don’t we directly compute the Bayes error? Estimating $P(x|1)$ and $P(x)$ itself has already been a difficult task. What’s more, in the experiments, the densities are estimated via 10-bin histograms, and I am afraid this may fail badly in slightly higher dimensional settings.
3) After deriving the conditional risk $C_{poly-n}$, we can pick a predicting function $f_\phi$ and work out a “better” loss function $\phi$ (using the method described in [1]). Then we minimize (8) and obtain an optimal $f_\phi^*$ right? In such a way, we optimized our predicting function with respect to a tighter upper-bound of Bayes error. If this algorithm is feasible, we can obtain an estimate of minimum risk $R_{C_\phi}$ without learning density $P(x)$.


4) In Theorem 7, it is stated that two distributions $P(u)$ and $Q(u)$ are two embedded Gaussian distributions. However, in practice, if the Gaussian kernel is used, the sample is embedded into an infinite-dimensional space and it is not clear how to define a Gaussian distribution in such a space. Similar issue appears in (62).

5) Is Section 6 still about two-sample test? How to justify the combination of different discrepancy measures? This section is more like a heuristic approach to me rather than a principled methodology.

Finally, If the coefficients of SVM (or Fisher Discriminant Analysis) are used for projecting the data onto the one-dimensional space, can we do two sample test without using the minimum risk? For example, is it possible to learn two density functions $P$ and $Q$ on projected samples and compare them using Kullback-Leibler divergence? Comparing this naive method, what are the advantages of the proposed method? Please clarify.