Reviewer #1:

Writing. We will fix the typos, notation inconsistencies (e.g. about the parameters ϵ and η), and incorporate the reviewer's formatting suggestions.

Hyperparameters. We will indicate the choice of hyperparameters in our experiments. Specifically, for $\mathrm{iForest}$, we used the default values suggested by the authors, namely sub-sampling size being 256 and number of trees being 100. For LODA, we used 100 projections with each projection using approximately $\sqrt{d}$ features.

Theorem 3.5. We thank the reviewer for checking the correctness and steps of our proofs, which we will expand to provide adequate details. In fact, Theorem 3.3 and 3.4 hold for most p in $[0, 1)$, except for values that are too close (in a quantifiable way) to 0 or 1. The results in Theorem 3.5 remain valid, and we will expand the arguments in our proof to describe the effect of boundary cases (which only impacts the constants). The proof for Equation (3) of Theorem 3.5 will follow easily from Theorem 3.4 after the update. We will also set $k = \lfloor np \rfloor$, throughout.

Comparison with refs. [1], [2] and [3]. We will comment on similarities and differences of our work with [1], [2] and [3], which are all essentially targeting the estimation of high-density points/regions. The reviewer has summarized them well. We would also add that we provide a new analysis of the minimal separation between the distribution of the normal and of the anomalous observation. (Finally, assumption A1 is eq. 24 in Theorem 12 of [14]).

Software. We apologize for the confusion. We will make our code available through a linked public github repository.

Reviewer #2:

Table 1 & 2. We will further emphasize the difference in interpretation between Table 1 and Table 2, and explain why we are using different evaluation metrics. Essentially, the failure rate serves as a metric for the average performance of different methods across the 20K synthetic datasets, whereas we believe practitioners may be more interested in obtaining raw AUC and AP scores on each real dataset.

Practical Interpretation. Figure 2 gives an example of a practical interpretation of Proposition 3.6: DTM will not make any mistake when the anomalies are sufficiently separated from the normal points, and identify the region for mis-classification when they are too close. We will include more illustration of the behavior of DTM in different scenarios, especially the ones where mistakes will be made near the boundary of the support of the distribution of the normal observations. We will also include additional commentary on the main theoretical results of Section 3.3.

More precise description of the simulation results. We will expand our comparisons of the different methods under various cases in our supplementary material. Each synthetic dataset is defined by a set of parameters (anomaly rate, difficulty, clusteredness, etc). We will indicate explicitly the performance of each method under each case, and comment on how the dataset parameters are related to the methods' performance.

Reviewer #3:

How realistic are our assumptions and Proposition 3.6. Our assumptions are fairly standard and generally regarded as mild in the literature on DTM and on geometric inference and are needed to rule out pathological cases. Yet, they capture a wide range of distributions. The assumptions in Proposition 3.6 encompass what are arguably some of the simplest instances of anomaly detection problems in fully non-parametric settings. We believe it is important to start with these cases to appreciate the difficulty of the task. We will add a remark.

Theory for Ensemble Methods. We are not aware of any theoretical analysis of ensemble methods, iForest and LODA in particular. Though this is a very important and broad question, it is also notoriously difficult and outside the scope of our paper.

Sensitivity Analysis. The only parameters for DTM are k and q. We set k to be the same as that for LOF for comparable results, and discussed the performance of DTM for $q = 1, 2, \infty$. We will include a discussion on sensitivity to the parameters and point the readers to the relevant literature.

Related Literature. We are thankful for the list of references, which we will definitely cite and comment on. We will mention the good performance of those sub-sampling and ensemble-based NN methods. It appears that the theoretical analyses from those references differ from ours significantly, in both assumptions and goals. We believe our theoretical results provide novel and complementary insights into the performance of KNN-based methods both for anomaly detection and for more general tasks.

Reviewer #4:

Overlapping Support. We agree: allowing for overlapping support is an interesting and more realistic case. However, to analyze it, it is necessary to add further assumptions on how the two distributions differ. In this paper, we consider a simpler case where we essentially leave the distribution of anomalous points unconstrained.

Boxplots. We agree: boxplots may be uninformative in this case. We used them as a coarse summary of the performance of 6 methods on 23 real datasets. Due to space limitations, we had to leave the full performance table (Table 3) in the supplementary file. We will add an extra row at the bottom of the table, indicating the average rank of each method, and conduct a pairwise Wilcoxon signed rank test.

Complexity of the datasets in the experiments. Please see response to Reviewer #2 point 3.

Connection between empirical and theoretical studies. We use section 2 as a motivation for the theoretical analysis in section 3, followed by a practical illustration of our theoretical results in section 3.4. We will be happy to hear and follow the reviewer suggestions on how to link these two parts seamlessly.