We greatly appreciate your efforts and precious time for providing us with helpful comments and fascinating ideas!

R1: Contribution is limited as this work is heavily based on Samworth (2012). Furthermore, Samworth (2012) suggests using cross-validation (CV); we may find good choices numerically without explicit formulas.

As written in line 92, we first note that the weights obtained via multiscale k-NN are DIFFERENT from Samworth (2012) as shown in Figure 2 in Supplement F, though they attain the same convergence rate (in terms of the order w.r.t. n). Although multiscale k-NN is heavily compared with Samworth (2012) as it is only one baseline in the same setting, these two methods are based on different ideas: Taylor-series of the risk is directly optimized in Samworth (2012) whereas multiscale k-NN considers a regression. Secondly, Samworth (2012) chooses only k by CV (with the weights $\boldsymbol{w}_k = (w_1, w_2, \dots, w_k)$, that can be explicitly obtained only for the limited cases $\beta \in \{2, 4\}$). Namely, equations should be solved to obtain the weights $\boldsymbol{w}_k$; the issue of Samworth (2012) remains, even if CV is utilized. Although we may conduct CV to choose $\boldsymbol{w}_k$ directly from $\mathcal{W}^k$ (for some set $\mathcal{W} \subset \mathbb{R}$ with $|\mathcal{W}| = m$), it requires the computational complexity $O(m^k)$, which is too large to compute in practice (as $k \geq 10^2$ in many practical cases). Therefore, multiscale k-NN still has advantages, compared to Samworth (2012) equipped with CV. As these are confusing points which are not well explained, we would like to revise the current manuscript to describe these advantages clearer.

R1: Is there some general strategy to set the parameters including V?

The multiscale k-NN attains the improved convergence rate for **any** combination of $V > \lfloor \beta/2 \rfloor + 1$ and $\ell_1 < \ell_2 < \cdots < \ell_V$, as the weights $\boldsymbol{w}_k$ are automatically adapted to the setting of $(V, \ell_1, \ell_2, \dots, \ell_V)$ via the regression. However, the smoothness β of the underlying function η cannot be obtained in practice; this is a common problem with the local polynomial (LP) regression. From both theoretical and application perspectives, we may simply employ a large V (e.g., $V = 100$) so that V is expected to be larger than $\lfloor \beta/2 \rfloor + 1$. Even if such a large V is employed, the computational complexity for the regression remains very small, as the number of regression coefficients to be estimated is only $1 + V$. It is different from the LP regression, as LP leverages $1 + d + d^2 + \cdots + d^V$ terms to attain the same convergence rate.

R1, R2: There is only a limited empirical analysis.

Although the main purpose of this paper is to provide an intuitive idea to understand how to obtain a faster convergence rate, we agree with this comment; we will add some experiments to emphasize the advantages of the improved rate.

R2: Whereas the proof may be of value of this paper, the authors state that it follows closely from Chaudhuri and Dasgupta (2014), and all details are relegated to the supplement.

The most important point is how to evaluate the asymptotic bias (and variance); our Theorem 1 indicates the order of the reduced bias of the multiscale k-NN, though its straightforward proof is mostly based on tedious Taylor expansion, which may not be deserving of explanation in detail in the main body. This proof is independent of Chaudhuri and Dasgupta (2014). Once the asymptotic bias (and variance) are obtained, following Chaudhuri and Dasgupta (2014) almost yields Theorem 2 (as well as many of nonparametric theories), though there remain some tedious technical issues to be considered. We will revise the current proof sketch so that the proof can be grasped more easily.

R2: The vast majority of the paper is dedicated to background review.

Although the detailed background review is necessary for explaining the position of this paper (among papers written with different notations/assumptions), we would like to revise the manuscript to condense the descriptions.

R2: It would benefit the paper greatly to provide some more discussion with the conditions in Theorem 2.

We agree with your comment. We would like to add some explanations (discussions) on the conditions. For instance, for (C-2): it is for selecting $k_1, k_2, ..., k_V$ so that the corresponding $r_1, r_2, ..., r_V$ distribute with regular intervals.

R3: The method seems to provide little benefit over LP estimators.

Whereas LP estimator considers the polynomial function $g(x)$ defined for $x \in \mathbb{R}^d$, multiscale k-NN considers $f(r)$ defined for the radius $r > 0$; the numbers of coefficients to be estimated are $1 + d + d^2 + \cdots + d^V$ for LP and $1 + V$ for multiscale k-NN for attaining the same convergence rate. Furthermore, multiscale k-NN is expected to inherit the favorable properties of the k-NN, such as the adaptability to the underlying manifold of the data vectors (Cheng and Wu, 2013). We are thinking about proving such favorable properties in future works.

References

Chaudhuri, K. and Dasgupta, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems 27*, pages 3437–3445. Curran Associates, Inc.

Cheng, M.-y. and Wu, H.-t. (2013). Local linear regression on manifolds and its geometric interpretation. *JASA*, 108(504):1421–1434.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, 40(5):2733–2763.