Statistical Optimality of Interpolated Nearest Neighbor Algorithms

Yue Xing
Department of Statistics
Purdue University
West Lafayette, Indiana, USA
xing49@purdue.edu

Qifan Song
Department of Statistics
Purdue University
West Lafayette, Indiana, USA
qfsong@purdue.edu

Guang Cheng
Department of Statistics
Purdue University
West Lafayette, Indiana, USA
chengg@purdue.edu

Abstract

In the era of deep learning, understanding over-fitting phenomenon becomes increasingly important. It is observed that carefully designed deep neural networks achieve small testing error even the training error is close to zero. One possible explanation is that for many modern machine learning algorithms, over-fitting can greatly reduces the estimation bias, while not increases the estimation variance too much. To illustrate the above idea, we prove that our interpolated nearest neighbors algorithm achieves the minimax optimal rate in both regression and classification regimes, and observe that they are empirically better than the traditional k nearest neighbor method in some cases.

1 Introduction

In deep learning, with the structure of neural networks getting more and more complicated, computer scientists proposed various approaches, such as Dropout, to handle the possible over-fitting issues (Srivastava et al. (2014); Gal and Ghahramani (2016); Chen and Lin (2014); Guo et al. (2016); LeCun et al. (2015)). However, recent studies, for example Zhang et al. (2016), demonstrate that deep neural networks have a small generalization error even when the training data is perfectly fitted. Similar phenomenon of strong generalization performance for over-fitted models occurs in other modern machine learning algorithms as well, including kernel machines, boosting and random forests.

Inspired by these observations, this work will investigate the statistical optimality of perfectly fitted (interpolated) models by nearest neighbors algorithms (NN). Specifically, given training data $\{x_i, y_i\}_{i=1}^n$, we study the regression estimator

$$\widehat{\eta}(x) = \sum_{x_i \in N_k(x)} w_i y_i,$$

and its corresponding classifier

$$\widehat{g}(x) = 1\{\widehat{\eta}(x) > 1/2\},$$

where $N_k(x)$ denotes the set of the nearest k neighbors of x, and the weight w_i's are designed in the way such that $\widehat{\eta}(x_i) = y_i$.

Traditional k-NN assigns $w_i = 1/k$ and does not perfectly fit the data. Its asymptotic behavior and convergence rate have been well studied by many works (Cover (1968); Wagner (1971); Fritz (1975); Schoenmakers et al. (2013); Sun et al. (2016); Chaudhuri and Dasgupta (2014)).

Belkin et al. (2018a,b) designed an interpolated NN algorithm by a normalized polynomial weight function, and further extended this idea to Nadaraya-Watson kernel regression. However, the results under their weighting scheme (Belkin et al., 2018a) are not proven to be optimal. And our goal here is to re-design a new type of weighting scheme, such that we can attain rate-optimality for both regression and classification objectives.

Our first contribution is to propose a new interpolated weighting scheme, under which both regression and classification estimates are minimax optimal. More specifically, the mean squared error (MSE) of $\widehat{\eta}(x)$ and the risk bound of the classifier $\widehat{g}(x)$ (under the margin condition of Tsybakov (2004)) both achieve minimax optimal rates.

Our second contribution is to provide an intuitive explanation on why the interpolated-NN can perform potentially better than traditional k-NN. In fact, there is a bias-variance trade-off for the nearest neighbor methods: traditional k-NN minimized the variance to some extent, while interpolated-NN tries to reduce the bias. Although theoretically both of them attain the same optimal rate of MSE, our empirical studies show that interpolated-NN always yields better estimation and predictions. We conjecture the superior performance of interpolated-NN over k-NN is due to a constant level difference of convergence speed, that is, the order of convergence rate for interpolated-NN and k-NN are same, but the multiplicative constant of the convergence rate is smaller for interpolated-NN.

2 Interpolated-NN Algorithm and Model Assumptions

Let $\mathcal{X} \subset \mathbb{R}^d$ be the support of X and μ be the probability measure of X on $\mathcal{X}$. Define $\eta(x) = E(Y|X = x)$, where the response variable Y can be either binary (classification problem) or continuous (regression problem). Given n iid observations $\{x_i, y_i = y(x_i)\}_{i=1}^n$ and a test sample x, let $x_{(i)}$ denote the ith nearest neighbor of x under $\mathcal{L}_2$ distance, and a weighted-NN algorithm predicts the mean response value as

$$\widehat{\eta}(x) = \sum_{i=1}^{k} w_i y(x_{(i)}),$$

where $w_i = w_i(x_1, \ldots, x_n)$ are some data-dependent nonnegative weights satisfying $\sum_{i=1}^k w_i = 1$.

To induce exact data interpolation, it is sufficient to require $w_i \to 1$ as $\|x_{(i)} - x\| \to 0$. Following Belkin et al. (2018a), we construct the weight as:

$$w_i \quad = \quad \frac{\phi\big(\frac{\|x_{(i)}-x\|}{\|x_{(k+1)}-x\|}\big)}{\sum_{j=1}^k \phi\big(\frac{\|x_{(j)}-x\|}{\|x_{(k+1)}-x\|}\big)},$$

for some positive function ϕ on [0,1] which satisfies $\lim_{t\to 0} \phi(t) = \infty$, and if $\|x_{(i)} - x\| = 0$, we define $w_i = 1$. The denominator term $\|x_{(k+1)} - x\|$ is for normalization purpose. Note that conditional on $X_{(k+1)}$, $\|X_i - x\|/\|X_{(k+1)} - x\|$s are independent variables in [0,1] for all X_i belonging to the k-neighborhood of x. The function ϕ plays a crucial role for the analysis of $\widehat{\eta}$ and $\widehat{g}$. For example, if $\phi(t)$ increases too fast as $t \to 0$, the weighted average $\widehat{\eta}$ is always dominated by $y(x_{(1)})$. In this case, the variance of $\widehat{\eta}$ and $\widehat{g}$ will be too large. Thus the following condition on the choice of ϕ is needed:

A.0 For any random variable $T \in [0, 1]$ whose density is bounded, the moment generating function of $\phi(T)$ exists, i.e., there exist some $s > 0$ and M, such that $\mathbb{E}(e^{s\phi(T)}) < M$.

Technically, this condition allows us to bound the un-normalized weight average $\sum_{i=1}^k \phi(\|x_{(i)} - x\|/\|x_{(k+1)} - x\|)y(x_{(i)})$ by exponential concentration inequalities. In general, any positive function ϕ satisfying $\phi(1/u) = O(\log(u))$ as $u \to \infty$ meets the above condition, and one typical example could be $\phi(t) = 1 - \log(t)$. Belkin et al. (2018a) chose ϕ to be $\phi(t) = t^{-\kappa}$ for some $0 < \kappa < d/2$, which unfortunately doesn't satisfy the above condition. Visual comparison among different choices

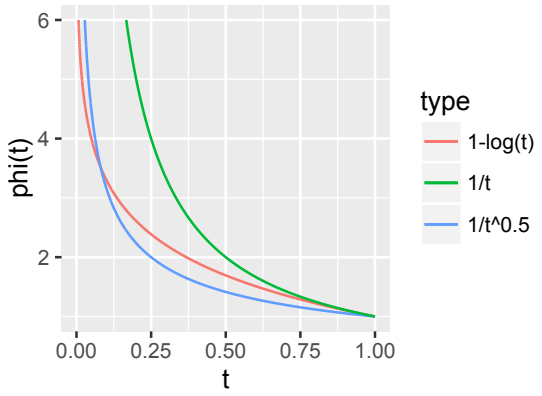

Figure 1: Interpolated Weighting Scheme

of ϕ can be found in Figure 1. Some toy regression examples are demonstrated in the supplementary material to compare different weighting schemes of interpolated-NN with k-NN.

Recall that the classifier is defined as $\widehat{g}(x) := 1\{\widehat{\eta}(x) \geq 1/2\}$, with the Bayes classifier $g(x) := 1\{\eta(x) \geq 1/2\}$. Define the point-wise excess risk $R_{n,k}(x) - R^*(x)$, where

$$\begin{aligned} R_{n,k}(x) &= P(Y \neq \widehat{g}(x)|X = x), \\ R^*(x) &= \min(\eta(x), 1 - \eta(x)). \end{aligned}$$

We next study the asymptotic rate of mean squared error of $\widehat{\eta}$ and expected excess risk for $\widehat{g}$ by imposing some regularity conditions. For any d-dimensional real-valued x in $\mathcal{X}$, let $B(x, r)$ represent the closed ball with radius r and center x.

A.1 Finite variance: $\bar{\sigma}^2 := \sup_{x \in \mathcal{X}} Var(Y|X = x) < \infty$.

A.2 Smoothness condition: $|\eta(x) - \eta(y)| \leq A\|x - y\|^\alpha$ for some $\alpha \in (0, 1]$.

A.3 Regularity condition: let λ be the Lebesgue measure on $\mathbb{R}^d$, then there exists positive (c_0, r_0) such that for any x in the support $\mathcal{X}$,

$$\lambda(\mathcal{X} \cap B(x, r)) \geq c_0 \lambda(B(x, r)),$$

for any $0 < r \leq r_0$.

A.4 The density of X is bounded and twice-continuously differentiable.

A.5 Margin condition: $P(|\eta(x) - 1/2| < t) \leq Bt^\beta$.

Assumption (A.2) is commonly assumed in the literature. A larger value of α implies more accurate estimation of η due to the minimax rate $O(n^{-2\alpha/(2\alpha+d)})$. Assumption (A.3) was first introduced by Audibert and Tsybakov (2007), and it essentially ensures that for any $x \in \mathcal{X}$, all its k nearest neighbors are sufficiently close to x with high probability. If $\mathcal{X} = \mathbb{R}^d$ or $\mathcal{X}$ is convex, this regularity condition is automatically satisfied. Assumption (A.5) is the so-called Tsybakov low noise condition (Audibert and Tsybakov (2007); Tsybakov (2004)), and this assumption is of interest only in the classification context. Under smoothness condition (A.2) and marginal condition (A.5), it is well known that the optimal rate for nonparametric regression and classification are $\mathbb{E}[(\widehat{\eta}(X) - \eta(X))^2] = O(n^{-2\alpha/(2\alpha+d)})$ and $\mathbb{E}[R_{n,p}(X) - R^*(X)] = O(n^{-\alpha(\beta+1)/(2\alpha+d)})$, respectively. Here, $\mathbb{E}$ means the expectation over all observed data $\{x_i, y_i\}_{i=1}^n$ and new observation $X \sim \mu$.

The analysis of classification is very subtle especially when $\eta(x)$ is near $1/2$. Hence, we need to have the following partition over the space $\mathcal{X}$. Define $v_p(x)$ as the p-th quantile of $\|X - x\|$, and

$$\eta(A) = \frac{1}{\mu(A)} \int_A \eta d\mu.$$

Also define

$$\mathcal{X}_{p,\Delta}^{+} = \{x \in \mathcal{X} | \eta(x) > \frac{1}{2}, \eta(B(x,r)) \geq \frac{1}{2} + \Delta, \forall r < v_p(x)\},$$
$$\mathcal{X}_{p,\Delta}^{-} = \{x \in \mathcal{X} | \eta(x) < \frac{1}{2}, \eta(B(x,r)) \leq \frac{1}{2} - \Delta, \forall r < v_p(x)\},$$

with the decision boundary area:

$$\partial_{p,\Delta} = \mathcal{X} \setminus (\mathcal{X}_{p,\Delta}^{+} \cup \mathcal{X}_{p,\Delta}^{-}).$$

3 Main Results

In this section, we will present our asymptotic results for the interpolated-NN algorithm described in Section 2. Our main Theorems 2 and 4 state that the interpolated estimator $\widehat{\eta}$ and classifier $\widehat{g}$ are asymptotically rate-optimal in terms of MSE and excess risk, respectively. In other words, data interpolation or data over-fitting doesn't necessarily jeopardize the statistical performance of a learning algorithm, at least in the minimax sense. These theoretical findings are supported by our numerical experiments in Section 4.

3.1 Over-Fitting Is Sometimes Even Better

Let us start from an intuitive comparison between the interpolated-NN and traditional k-NN. For any weighted-NN algorithm, the following bias-variance decomposition (Belkin et al. (2018a)) holds (with α and A in Assumption (A.2)):

$$\mathbb{E}\big\{(\widehat{\eta}(x) - \eta(x))^2\big\} \leq A^2 \mathbb{E}\bigg[\sum_{i=1}^{k} W_i \|X_{(i)} - x\|^{\alpha}\bigg]^2 + \sum_{i=1}^{k} \mathbb{E}\bigg[W_i^2 (Y(X_{(i)}) - \eta(X_{(i)}))^2\bigg], \quad (1)$$

where the first and second term represents squared bias and variance, respectively, and W_i denotes the random weight $w_i(X_1, \ldots, X_n)$. The k-NN ($W_i \equiv 1/k$) can be interpreted as optimal weight choice which minimizes the variance term in the case that $\mathrm{Var}(Y_i)$ are constant, e.g., under regression setting that $Y_i = \eta(X_i) + \epsilon_i$ with iid error term ϵ_i, or under classification setting that $Y_i \sim \mathrm{Bern}(\eta(X_i))$ with constant function η. On the other hand, interpolated-NN assigns larger weight for closer neighbor, which will result in a smaller value for the weighted average $\sum_{i=1}^{k} W_i \|X_{(i)} - x\|^{\alpha}$, i.e., the bias term gets smaller. Therefore, we argue that k-NN and interpolated-NN employ different strategies in reducing the MSE. The former one emphasizes reducing the variance, while the latter one emphasizes reducing the bias. The above intuitive arguments are also well validated by our toy examples described in the supplementary material.

Later, we will prove that the proposed interpolated-NN method, although potentially enlarge the variance term, is still minimax rate-optimal for both regression and classification cases.

3.2 Regression

As seen in the decomposition (1), the bias term involves $\|X_{(i)} - x\|$, which is the empirical quantile for $\|X - x\|$. Hence our first lemma, whose proof can be found in the supplementary material, will study the asymptotic behavior of this empirical quantile.

Let $F_x(\cdot)$, $\widehat{F}_x(\cdot)$ and $f_x(\cdot)$ be the c.d.f., empirical c.d.f and p.d.f. of $\|X - x\|$, $v_p(x)$ and $\widehat{v}_p(x)$ be the pth quantile and empirical quantile of $\|X - x\|$.

Lemma 1. *Given any $x \in \mathcal{X}$, let $p = k/n$ (thus $\widehat{v}_p(x) = \|X_{(k)} - x\|$), and*

$$A_2(x) = F_x(\widehat{v}_p(x)) + \widehat{F}_x(v_p(x)) - \widehat{F}_x(\widehat{v}_p(x)) - F_x(v_p(x)).$$

Under Assumption (A.4), if $|\widehat{v}_p(x) - v_p(x)|$ and p are much larger than $O(1/n)$, and $v_p(x) = O(p^{1/d})$ with $p^{1/d-1/2}$ much larger than $O(n^{-1/2})$, $f(v_p(x)) = O(p^{1-1/d})$, then

$$\mathbb{E}A_2^2(x) \to O(p^{1/2}/n^{3/2}).$$

The above lemma, combined with technical tool developed in Sun and Hong (2009, 2010), facilitates our theoretical analysis for the bias term in (1). As for the variance term in (1), we can bound it by the similar approach used by Belkin et al. (2018a). Together, it leads to the following theorem.

Theorem 2 (Rate-Optimality of Interpolated-NN Regression)**.** *Under Assumption (A.0)-(A.5), for any fixed $x \in \mathcal{X}$, there exists constants $c_1 \to 1^+$, (μ, x)-dependent c_2, when $n \to \infty$,*

$$\mathbb{E}\big\{(\widehat{\eta}(x) - \eta(x))^2\big\} \leq A^2 c_1 \left(\frac{k}{n}\right)^{2\alpha/d} + c_2 \bar{\sigma}^2 \left(k e^{-k/4} + \frac{1}{k}\right). \tag{2}$$

Therefore, taking $k \asymp (n^{2\alpha/(2\alpha+d)})$, $\mathbb{E}\big\{(\widehat{\eta}(x) - \eta(x))^2\big\}$ reaches the optimal rate of $O(n^{-2\alpha/(2\alpha+d)})$.

Proof. First, it is trivial to see that the bias-variance representation (1) can be further bounded by

$$\mathbb{E}\big[\widehat{\eta}(x) - \eta(x)\big]^2 \leq A^2 \mathbb{E}\|X_{(k)} - x\|^{2\alpha} + \bar{\sigma}^2 k \mathbb{E}(W_1^2).$$

Therefore the remaining task is to figure out the convergence rate of $\mathbb{E}\|X_{(k)} - x\|^{2\alpha}$ and calculate $\mathbb{E}(W_1^2)$.

For bias term $\mathbb{E}\|X_{(k)} - x\|^{2\alpha}$, note that $\|X_{(k)} - x\|$ is the p(=k/n)th empirical quantile of F_x, i.e., $\widehat{v}_p(x) = \|X_{(k)} - x\|$ (for simplicity, we write F_x as F, f_x as f, $v_p(x)$ as v_p and $\widehat{v}_p(x)$ as $\widehat{v}_p$ in what follows), and this quantile is called Value-at-Risk in the area of finance. From Sun and Hong (2009), we have

$$\widehat{v}_p \;=\; v_p + \frac{1}{f(v_p)}\left(p - \frac{1}{n}\sum_{i=1}^{n} 1_{\{r_i \leq v_p\}}\right) + \frac{1}{f(v_p)}(A_1 + A_2 + A_3),$$

where $A_1 = O(f'(v_p)(\widehat{v}_p - v_p)^2)$, $A_2 = F(\widehat{v}_p) + \widehat{F}(v_p) - \widehat{F}(\widehat{v}_p) - F(v_p)$, and $A_3 = \widehat{F}(\widehat{v}_p) - F(v_p)$. Compared with $\widehat{v}_p - v_p$, $A_1/f(v_p)$ is a smaller order term. Since $p = k/n$, we always have $\widehat{F}(\widehat{v}_p) = F(v_p) = p$, hence $A_3 \equiv 0$.

Within some neighborhood of x, $B(x, r)$ with $r \leq r_0$, both maximum and minimum density of X are bounded. Denote them as $p_{\max}$ and $p_{\min}$. As $n \to \infty$, $v_p \leq r$, thus by Assumption (A.4), $v_p^d p_{\max} \geq k/n$ and $\mathrm{c} v_p^d p_{\min} \leq k/n = F(v_p)$ for some constant c. This implies that $v_p \geq (k/n p_{\max})^{1/d}$, and

$$f(v_p) \;=\; O\left(\frac{k}{n}\right)^{1-1/d}.$$

As a result, when $\alpha \in (0, 1]$, define $r_i = \|X_i - x\|$, we have

$$\mathbb{E}\left(p - \frac{1}{n}\sum_{i=1}^{n} 1_{\{r_i \leq v_p\}}\right)^{2\alpha} \leq \left[\mathbb{E}\left(p - \frac{1}{n}\sum_{i=1}^{n} 1_{\{r_i \leq v_p\}}\right)^2\right]^{\alpha} = \left[\frac{k}{n^2}\left(1 - \frac{k}{n}\right)\right]^{\alpha} \leq \left(\frac{k}{n^2}\right)^{\alpha}.$$

Since $k \leq n$,

$$\frac{1}{f(v_p)^{2\alpha}}\mathbb{E}\left(p - \frac{1}{n}\sum_{i=1}^{n} 1_{\{r_i \leq v_p\}}\right)^{2\alpha} = O\left(\frac{k}{n}\right)^{2\alpha(1/d-1)} \mathbb{E}\left(p - \frac{1}{n}\sum_{i=1}^{n} 1_{\{r_i \leq v_p\}}\right)^{2\alpha} = o\left(\frac{k}{n}\right)^{2\alpha/d}.$$

In terms of A_2, based on Assumption (A.4) and Lemma 1, for $\alpha \in (0, 1]$,

$$\begin{aligned}
\frac{1}{f(v_p)^{2\alpha}}\mathbb{E}A_2^{2\alpha} \;&=\; O\left(\frac{k}{n}\right)^{2\alpha(1/d-1)} \mathbb{E}A_2^{2\alpha} \\
&=\; O\left(\frac{p^{-2\alpha+\alpha/2}}{n^{3\alpha/2}}\right) O\left(\frac{k}{n}\right)^{2\alpha/d} \\
&=\; O\left(\frac{p^{-3\alpha/2}}{n^{3\alpha/2}}\right) O\left(\frac{k}{n}\right)^{2\alpha/d}.
\end{aligned}$$

As a result, when p is larger than $O(n^{-1})$, which is satisfied when taking $k \asymp (n^{2\alpha/(2\alpha+d)})$, we have

$$O\left(\frac{p^{-3\alpha/2}}{n^{3\alpha/2}}\right) = o(1),$$

hence,

$$A^2\mathbb{E}\|X_{(k)} - x\|^{2\alpha} \leq A^2 c_1 \left(\frac{k}{n}\right)^{2\alpha/d},$$

with $c_1 \to 1^+$ in n.

For $\mathbb{E}W_1^2$, similar with Belkin et al. (2018a), we can bound it as

$$\begin{aligned}\mathbb{E}W_1^2 \quad \leq \quad & P(\|X_{(k+1)} - x\| > v_{2p}) \\ & + \mathbb{E}[W_1^2 | \|X_{(k+1)} - x\| < v_{2p}],\end{aligned}$$

where v_{2p} is the $2k/n$th quantile of $\|X - x\|$.

By the same argument in Claim A.6 of Belkin et al. (2018a), the probability $P(\|X_{(k+1)} - x\| > v_{2p})$ under Assumption (A.3) and (A.4) can be bounded by $O(\exp(-k/4))$.

On the other hand, by the arguments in Lemma 10 of Chaudhuri and Dasgupta (2014), conditional on $\|X_{(k+1)} - x\|$, for those X_i's that belong to x's k-neighborhood, $\|X_i - x\|/\|X_{(k+1)-x}\|$ are iid random variables in [0,1] . It follows that

$$\begin{aligned}\mathbb{E}[W_1^2 | \|X_{(k+1)} - x\| < v_{2p}] =& \mathbb{E}\left[\frac{\left[\phi\left(\frac{\|X_1 - x\|}{\|X_{(k+1)} - x\|}\right)\right]^2}{\left[\sum_{i=1}^k \phi\left(\frac{\|X_i - x\|}{\|X_{(k+1)} - x\|}\right)\right]^2} \middle| \|X_{(k+1)} - x\| < v_{2p} \right] \\ \leq & \frac{1}{k^2}\mathbb{E}\left[\phi\left(\frac{\|X_1 - x\|}{\|X_{(k+1)} - x\|}\right)^2 \middle| \|X_{(k+1)} - x\| < v_{2p}\right].\end{aligned}$$

An $O(1/k^2)$ result can be directly obtained through the existence of $\mathbb{E}e^{\phi}$ (Assumption (A.0)). □

Theorem 2 proves the point-wise MSE convergence result for $\widehat{\eta}$. By assuming that $\mathcal{X}$ is compact (which is also assumed in Belkin et al. (2018a) and Chaudhuri and Dasgupta (2014)), it is not difficult to see that there exist constant c_1 and c_2 in (2) such that uniformly holds for all $x \in \mathcal{X}$.

Corollary 3. *Under Assumption (A.0)-(A.4), if $\mathcal{X}$ is compact, then there exists constants $c_1 \to 1^+$, μ-dependent c_2, such that when $n \to \infty$,*

$$\mathbb{E}\big\{(\widehat{\eta}(X) - \eta(X))^2\big\} \leq A^2 c_1 \left(\frac{k}{n}\right)^{2\alpha/d} + c_2\bar{\sigma}^2\left(ke^{-k/4} + \frac{1}{k}\right).$$

Therefore, taking $k \asymp (n^{2\alpha/(2\alpha+d)})$, $\mathbb{E}\big\{(\widehat{\eta}(X) - \eta(X))^2\big\}$ reaches the optimal rate of $O(n^{-2\alpha/(2\alpha+d)})$.

Remark 1. *It is worth to mention that Assumption (A.0) is not necessary for Theorem 2 and Corollary 3. Our proof can be easily adopted to show that interpolated-NN regression estimator with polynomial $\phi(t) = t^{-\kappa}$ is also rate-optimal.*[1]

Remark 2. *Instead of nearest neighbors method, one can alternatively consider kernel regression estimator $\sum_{\|x_i - x\| \leq h} w_i y(x_i)$ with interpolated weights (Belkin et al. (2018b)). In such a case, the theoretical investigation of pth quantile of F_x can be avoided, and if the bandwidth h is of the same order of v_p, one can still obtain the optimal rate.*

[1] Belkin et al. (2018a) claims the optimal rate of convergence of $\widehat{\eta}(x)$ with $\phi(t) = t^{-\kappa}$ based on heuristic argument on the order of $\mathbb{E}\|X_{(k)} - x\|$.

3.3 Classification

In this section, we investigate the theoretical properties of the interpolated-NN classifier $\widehat{g}$, and the next theorem establishes the statistical optimality of $\widehat{g}$ in terms of excess risk.

Theorem 4 (Rate-Optimality of Interpolated-NN Classification)**.** *Under Assumption (A.0)-(A.5), assume $\mathcal{X}$ is compact, then for any $\delta \in (0,1)$, taking $k \asymp (n^{2\alpha/(2\alpha+d)}(\log \delta)^{d/(2\alpha+d)})$, we have*

$$P(g(X) \neq \widehat{g}(X)) \leq \delta + C_0 \left(\frac{\log(1/\delta)}{n} \right)^{\frac{\alpha\beta}{2\alpha+d}} \tag{3}$$

for some constant $C_0 > 0$.

Moreover, taking $k \asymp (n^{2\alpha/(2\alpha+d)})$,

$$\mathbb{E}(R_{n,p}(X) - R^*(X)) \leq O(n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}). \tag{4}$$

We remark that the above bounds in (3) and (4) are the same as those in Chaudhuri and Dasgupta (2014).

Proof. Let $p = 2k/n$, following Belkin et al. (2018a); Chaudhuri and Dasgupta (2014), we obtain that for any fixed $\Delta \in (0,1)$,

$$\begin{aligned}
P(g(X) \neq \widehat{g}(X)) \quad \leq \quad & \mu(\partial_{2p,\Delta}) + P(E) \\
& + P\left(\sum_{i=1}^{k} W_i (Y(X_{(i)}) - 1/2) > 0 \middle| X \in \mathcal{X}^{-}_{2p,\Delta} \text{and} E^c \right) \\
& + P\left(\sum_{i=1}^{k} W_i (Y(X_{(i)}) - 1/2) < 0 \middle| X \in \mathcal{X}^{+}_{2p,\Delta} \text{and} E^c \right) \\
\leq \quad & B\left(\Delta + AC_1 \left(\frac{k}{n} \right)^{\alpha/d} \right)^{\beta} + B\exp(-k/4) \\
& + P\left(\sum_{i=1}^{k} W_i (Y(X_{(i)}) - 1/2) > 0 \middle| X \in \mathcal{X}^{-}_{2p,\Delta} \text{and} E^c \right) \\
& + P\left(\sum_{i=1}^{k} W_i (Y(X_{(i)}) - 1/2) < 0 \middle| X \in \mathcal{X}^{+}_{2p,\Delta} \text{and} E^c \right),
\end{aligned}$$

holds for some constants A, B and C_1, where $r_i = \|X_{(i)} - X\|$, $\mathrm{E} = \{r_{k+1} > v_{2p}(X)\}$. Denote $Z_i(x) = \phi(\|X_i - x\|/r_{k+1})(Y(X_i) - 1/2)$ for those k nearest X_i's. By Assumption (A.0) of ϕ and the arguments in Theorem 5 of Chaudhuri and Dasgupta (2014), given any fixed x, conditional on r_{k+1}, $Z_i(x)$ are iid sub-exponential variables. Hence by Bernstein inequality and compact $\mathcal{X}$ assumption, for some $C_2, C_3, C_4 > 0$, conditional on $X \in \mathcal{X}^{-}_{2p,\Delta}$ and E^c,

$$\begin{aligned}
P\left(\sum_{i=1}^{k} W_i (Y(X_{(i)}) - 1/2) > 0 \right) =& P\left(\sum_{i=1}^{k} Z_i - \mathbb{E}Z_i > -k\mathbb{E}Z_1 \right) \\
\leq& C_2 \exp\left(- C_3 k (\mathbb{E}Z_1)^2 \right) \\
\leq& C_2 \exp(-C_4 k \Delta^2).
\end{aligned} \tag{5}$$

Let $\delta := 2C_2 \exp(-C_4 k \Delta^2)$, then replacing Δ with δ, we have

$$P(g(X) \neq \widehat{g}(X)) \leq O\left(\sqrt{\frac{\log 1/\delta}{k}} + \left(\frac{k}{n} \right)^{\alpha/d} \right)^{\beta} + B\exp(-k/4) + \delta,$$

then the proof of (3) is completed by taking $k \asymp (n^{2\alpha/(2\alpha+d)}(\log \delta)^{d/(2\alpha+d)})$.

To prove (4), we follow the proof of Chaudhuri and Dasgupta (2014). Without loss of generality assume $\eta(x) < 1/2$. Define

$$\begin{aligned} \Delta_0 &= v_{2p}^{\alpha} = O(k/n)^{\alpha/d}, \\ \Delta(x) &= |\eta(x) - 1/2|, \end{aligned}$$

so that $\eta(B(x,r)) \leq \Delta_0 + \eta(x) = 1/2 - (\Delta(x) - \Delta_0)$ for $r < v_{2p}$, and x is not in $\partial_{p,\Delta(x)-\Delta_0}$.

From the definition of $R_{n,k}$ and R^*, for some constant c_1, with $\Delta(x) > \Delta_0$, we have

$$\begin{aligned} \mathbb{E}R_{n,k}(x) - R^*(x) &\leq 2\Delta(x)\Bigg[P(r_{(k+1)} > v_{2p}) \\ &\quad + P\bigg(\sum_{i=1}^{k} W_i(Y(X_{(i)}) - \eta(X_{(i)})) > \Delta(x) - \Delta_0\bigg)\Bigg] \\ &\leq \exp(-k/8) + 4\Delta(x)\exp(-c_1 k(\Delta(x) - \Delta_0)^2), \end{aligned}$$

where the last line is obtained after adopting Bernstein inequality (similarly to (5)).

Define $\Delta_i = 2^i\Delta_0$, for any $i_0 = 1, ...$, the excess risk can be bounded by

$$\begin{aligned} \mathbb{E}R_{n,k}(X) - R^*(X) &= \mathbb{E}(R_{n,k}(X) - R^*(X))1_{\{\Delta(X)\leq\Delta_{i_0}\}} \\ &\quad + \mathbb{E}(R_{n,k}(X) - R^*(X))1_{\{\Delta(X)>\Delta_{i_0}\}} \\ &\leq \mathbb{E}(2\Delta(X)1_{\{\Delta(X)\leq\Delta_{i_0}\}}) + \exp(-k/8) \\ &\quad + 4\mathbb{E}[\Delta(x)\exp(-k(\Delta(x) - \Delta_0)^2)1_{\{\Delta(X)>\Delta_{i_0}\}}] \\ &\leq 2\Delta_{i_0}P(\Delta(X) \leq \Delta_{i_0}) + \exp(-k/8) \\ &\quad + 4\mathbb{E}[\Delta(x)\exp(-k(\Delta(x) - \Delta_0)^2)1_{\{\Delta(X)>\Delta_{i_0}\}}], \end{aligned}$$

while

$$\begin{aligned} &\mathbb{E}[\Delta(x)\exp(-k(\Delta(x) - \Delta_0)^2)1_{\{\Delta_i<\Delta(X)\leq\Delta_{i+1}\}}] \\ \leq\ & \Delta_{i+1}\exp(-c_1 k(\Delta_i - \Delta_0)^2)P(\Delta(X) \leq \Delta_{i+1}) \\ \leq\ & c_2\Delta_{i+1}^{\beta+1}\exp(-c_1 k(\Delta_i - \Delta_0)^2). \end{aligned}$$

Taking

$$i_0 = \max\left(1, \left\lceil \log_2 \sqrt{\frac{2\beta+4}{k\Delta_0^2}} \right\rceil\right),$$

then for $i > i_0$, we have

$$\frac{\Delta_{i+1}^{\beta+1}\exp(-c_1 k(\Delta_i - \Delta_0)^2)}{\Delta_i^{\beta+1}\exp(-c_1 k(\Delta_{i-1} - \Delta_0)^2)} \leq 1/2. \tag{6}$$

Therefore, the sum of the excess risk for $i > i_0$ can be bounded, where

$$\begin{aligned} \mathbb{E}[\Delta(x)\exp(-k(\Delta(x) - \Delta_0)^2)1_{\{\Delta(X)>\Delta_{i_0}\}}] &\leq \sum_{i\geq i_0} c_2\Delta_{i+1}^{\beta+1}\exp(-c_1 k(\Delta_i - \Delta_0)^2) \\ &\leq O(\Delta_{i_0}^{\beta+1}). \end{aligned}$$

Finally, recall the definition of i_0, we have

$$\mathbb{E}R_{n,k}(X) - R^*(X) \leq \max\left(O\left(\frac{k}{n}\right)^{\frac{\alpha(\beta+1)}{d}}, O\left(\frac{1}{k}\right)^{\frac{\beta+1}{2}}\right).$$

The proof is completed after taking $k \asymp (n^{2\alpha/(2\alpha+d)})$. □

Remark 3. *It is worth to mention that, unlike Theorem 2 and Corollary 3 which are asymptotic results, Theorem 4 presents an non-asymptotic bound for misclassification rate and excessive risk.*

Remark 4. *Comparing with the choice $\phi(t) = t^{-\kappa}$ proposed by Belkin et al. (2018a), our choice (e.g., $\phi(t) = 1 - \log(t)$) leads to a sharp bound for classification error rate. Technically, this is due to the fact that slowly increasing $\phi(\cdot)$ allows us to use exponential concentration inequality.*

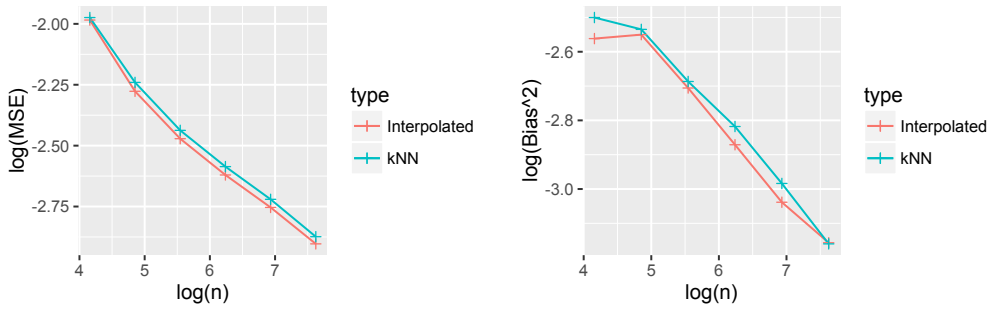

Figure 2: Optimal MSE and the Corresponding Bias for Regression Model 1

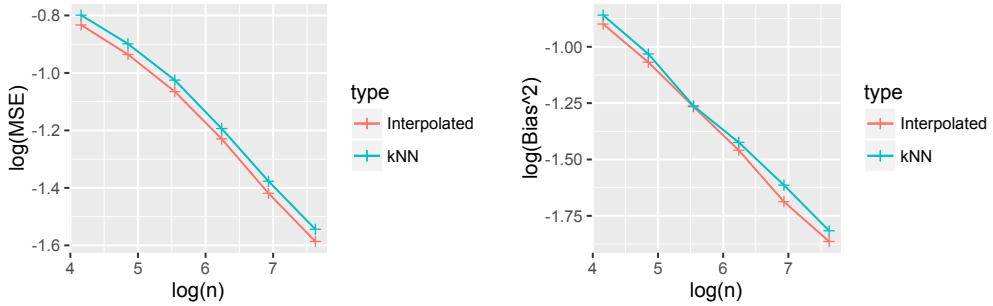

Figure 3: Optimal MSE and the Corresponding Bias for Regression Model 2

4 Numerical Experiments

In the section, we demonstrate several numerical studies to compare the performance of interpolated-NN with $\phi(t) = 1 - \log(t)$ and traditional k-NN, for both regression and classification problems.

4.1 Regression

We have two simulation setups. In the first setup, X follows Unif$[-3, 3]^{10}$. For y, we have

$$y = \psi(x) + \epsilon, \tag{7}$$

$$\psi(x) = \frac{N(x, 1, I_{10})}{N(x, 1, I_{10}) + N(x, 0, I_{10})}, \tag{8}$$

where $\epsilon \sim t(\text{df} = 5)$ and $N(x, \mu, \sigma)$ represents the density function of $N(\mu, \sigma)$ at x.

For each pair of (k, n), we sampled 1000 testing data points to estimate MSE, and repeated 30 times. For each n, we tried $k = 1, ..., n/2$. Based on the average of the 30 repetitions, we selected the minimum average MSE over the choices of k and recorded its corresponding square bias. From Figure 2, for both MSE and bias, the interpolated-NN and k-NN share the same rate (the decreasing slopes for both are similar), but the former is constantly better.

In the second experiment, X follows $N(0, I_5)$. The response y is defined as

$$y = (\sum_{i=1}^{5} x_i)^2 + \epsilon,$$

with $\epsilon \sim N(0, 1)$. The phenomenon demonstrated in Figure 3 is similar as that in Figure 2, although the support of X is not compact in this setup.

It is worth to mention, in both simulation, the optimal k values selected by interpolated-NN and k-NN doesn't have much difference.

4.2 Classification

For the simulation setup of classification problem, we consider that the two classes follow $N(0, I_5)$ and $N(\gamma \mathbf{1}, I_5)$ with γ = 0.1, 0.2, 0.5, 0.7, 1.0, 1.5. Training and testing samples are both generated from the mixture of these two classes with equal probability. Such an equal probability mixture represents the worst (most difficult) scenario. For each pair of n and γ, we tried $k = 1, ..., n/2$ for 30 times with 1000 test samples, and use the excess risk to determine the optimal k. From Figure 4, we observe that the interpolated-NN always has a smaller classification error (with a similar trend, though) than k-NN in most pairs of n and γ. Moreover, we record the optimal k for each n. As shown in Figure 5, interpolated-NN and k-NN have a similar pattern on the choice of optimal k across different settings of n and γ.

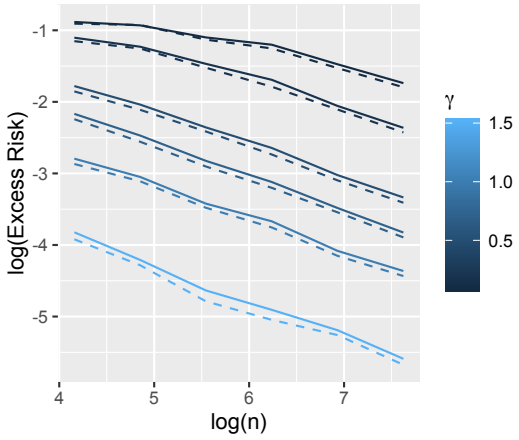

Figure 4: Optimal Excess Risk for Classification Model, Solid Line for k-NN, Dashed Line for interpolated-NN

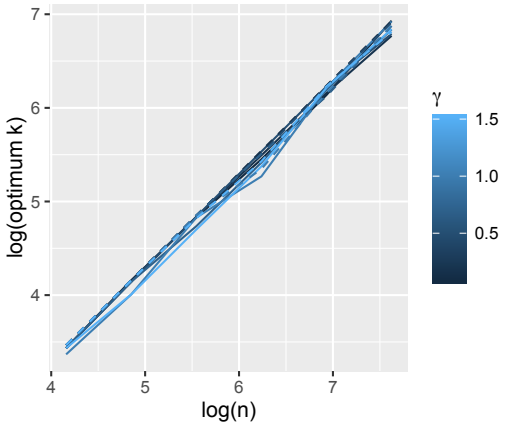

Figure 5: Optimal k for Classification Model

Remark 5. *The simulation results presented in above two sections clearly show that interpolated-NN and k-NN share the same order of convergence rate, as their decreasing treads are parallel. However, the interpolated-NN is always better than the k-NN. This observation strongly suggests that the convergence rate of interpolated-NN has a smaller multiplicative constant than the k-NN.*

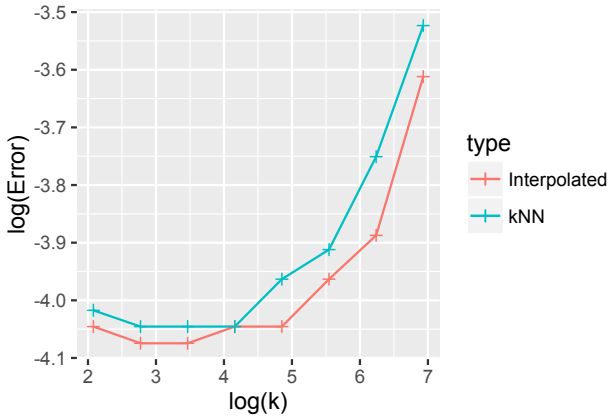

Figure 6: Performance in HTRU2

4.3 Real Data

In this section, we examine the empirical performance of the interpolated-NN in real data. The data HTRU2 is from Lyon et al. (2016) with sample size around $18,000$ and 8 continuous attributes. The dataset is first normalized, and 2,000 randomly selected samples are reserved for testing.

Instead of comparing regression performance through MSE, we compare the classification error on testing data in this real data example since the true Bayes risk is unknown. It can be seen from Figure 6 that the interpolated-NN is always better than k-NN regardless of the choice of k.

5 Conclusion and Discussion

In this paper, we firstly provide some insights on why sometimes an over-fitting weighting scheme can even beat traditional k-NN: the interpolated weights greatly reduce the estimation bias comparing with tradition k-NN. We then show the optimal convergence rates of interpolated-NN for both regression and classification. Even though the weighting scheme cause over-fitting, interpolated-NN can still obtain the optimal rate, as long as the weights are carefully designed.

In the end, we would like to point a few promising future directions: firstly, as we mentioned, in most cases of our simulations, the interpolated-NN performs better than k-NN. This motivates us to study the sharp convergence rate of our interpolated-NN, and we conjecture that interpolated-NN will have a smaller multiplicative constant in convergence rate.

Our simulation indicates that the interpolated-NN performs very well even when the compactness assumption of $\mathcal{X}$ is violated. Therefore, it will be of interests to extend our current work to unbounded $\mathcal{X}$ case. Especially, we notice that in Döring et al. (2017), the compact assumption on $\mathcal{X}$ can be relaxed for traditional k-NN algorithm. Similar results may hold for interpolated-NN as well.

In addition, our results are based on $\mathcal{L}_2$ metric. It is not hard to generalize it to $\mathcal{L}_p$, but it remains to study whether the statistical optimality is still valid for some more general metric spaces, such as Riemannian manifolds.

Finally, in the era of big data, a direct implementation of nearest neighbor algorithm on the whole dataset requires vast memory space and computing time. An embarrassingly parallel nearest neighbor algorithm, $\widehat{\eta}(x) = \sum_{j=1}^{s} \widehat{\eta}_j(x)/s$ where $\eta_j(x)$ is the nearest neighbor estimator based on a batch of data, is more computational attractive. Thus, it is of substantial interests to study how to design a rate-optimal interpolated weighting scheme for distributed-NN algorithm.

References

Audibert, J.-Y. and Tsybakov, A. B. (2007), "Fast learning rates for plug-in classifiers," *The Annals of statistics*, 35, 608–633.

Belkin, M., Hsu, D., and Mitra, P. (2018a), "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," *arXiv preprint arXiv:1806.05161*.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018b), "Does data interpolation contradict statistical optimality?" *arXiv preprint arXiv:1806.09471*.

Chaudhuri, K. and Dasgupta, S. (2014), "Rates of convergence for nearest neighbor classification," in *Advances in Neural Information Processing Systems*, pp. 3437–3445.

Chen, X.-W. and Lin, X. (2014), "Big data deep learning: challenges and perspectives," *IEEE access*, 2, 514–525.

Cover, T. M. (1968), "Rates of convergence for nearest neighbor procedures," in *Proceedings of the Hawaii International Conference on Systems Sciences*, pp. 413–415.

Döring, M., Györfi, L., and Walk, H. (2017), "Rate of convergence of k-nearest-neighbor classification rule," *The Journal of Machine Learning Research*, 18, 8485–8500.

Fritz, J. (1975), "Distribution-free exponential error bound for nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, 21, 552–557.

Gal, Y. and Ghahramani, Z. (2016), "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016), "Deep learning for visual understanding: A review," *Neurocomputing*, 187, 27–48.

LeCun, Y., Bengio, Y., and Hinton, G. (2015), "Deep learning," *nature*, 521, 436.

Lyon, R. J., Stappers, B., Cooper, S., Brooke, J., and Knowles, J. (2016), "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach," *Monthly Notices of the Royal Astronomical Society*, 459, 1104–1123.

Schoenmakers, J., Zhang, J., and Huang, J. (2013), "Optimal dual martingales, their analysis, and application to new algorithms for Bermudan products," *SIAM Journal on Financial Mathematics*, 4, 86–116.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014), "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, 15, 1929–1958.

Sun, L. and Hong, L. J. (2009), "A general framework of importance sampling for value-at-risk and conditional value-at-risk," in *Winter Simulation Conference*, pp. 415–422.

— (2010), "Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk," *Operations Research Letters*, 38, 246–251.

Sun, W. W., Qiao, X., and Cheng, G. (2016), "Stabilized nearest neighbor classifier and its statistical properties," *Journal of the American Statistical Association*, 111, 1254–1265.

Tsybakov, A. B. (2004), "Optimal aggregation of classifiers in statistical learning," *Annals of Statistics*, 135–166.

Wagner, T. (1971), "Convergence of the nearest neighbor rule," *IEEE Transactions on Information Theory*, 17, 566–571.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016), "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*.

A Toy Examples

Several toy examples are conducted to demonstrate nearest neighbor algorithm with interpolated weights.

In these examples, we take 30 one dimensional training samples $x_i = -5, -4, ..., 25$, and generate three choices of response (1) $y = 0 * x + \epsilon$, (2) $y = x^2 + 0 * \epsilon$, and (3) $y = (x-10)^2/8 + 5 * \epsilon$ where $\epsilon \sim N(0,1)$. In other words, the mean function $\eta(x)$ are (1) $\eta(x) \equiv 0$, (2) $\eta(x) = x^2$, and (3) $\eta(x) = (x-10)^2/8$. The number of neighbors k equals 10.

Three different weighting schemes are considered (1) $\phi(t) = 1 - \log(t)$, (2) $\phi(t) = 1/k$, and (3) $\phi(t) = t^{-1}$. Note that the first and third choices are interpolated weight, and the second choice is simply the traditional k-NN. In particular, the first $\phi(\cdot)$ satisfies Assumption (A.0). The four line "true" refers to $\eta(x)$.

Based on this setting, we plot the regression estimator $\widehat{\eta}(x)$ in Figure 7. In order to remove the boundary effect, we only plot the $\widehat{\eta}(x)$ within range $(0, 20)$. Note that for the second case, in order to make the difference more clear and recognizable, we only plot $\widehat{\eta}(\cdot)$ for x between 10 and 15.

There are several insights we can obtain from the results of this toy example.

First of all, interpolated weight does ensure data interpolation. As x gets closer to the some observed x_i, the estimator $\widehat{\eta}(x)$ is forced towards y_i. As a consequence, $\widehat{\eta}(x)$ is spiky for interpolated-NN. On contrast, the k-NN estimator is much more smooth.

Secondly, different weighting schemes lead to different balance between bias and variance of $\widehat{\eta}$. In the first setting that $\eta \equiv 0$, any weight-NN algorithm is unbiased, hence it corresponds to the extreme situation that bias is 0; the second model is noiseless, hence corresponds to the opposite extreme situation that variance is 0. In the no-bias setting, k-NN performs the best, and interpolated-NN estimators fluctuate a lot. In the noiseless case, k-NN has the largest bias, and $\phi(t) = t^{-1}$ leads to smallest bias. These observations are consistent to our arguments in Section 3.1 in the main text, i.e., k-NN tries to minimize the variance of nearest neighbor estimator as much as possible, while interpolated-NN tries to minimize the estimation bias. For the comparison between different interpolated weighting scheme, we comment that the faster $\phi(t)$ increases to infinite as $t \to 0$, the smaller bias it will yield. Thus $\phi(t) = t^{-1}$ leads to smaller bias then $\phi(t) = 1 - \log(t)$, at the expense of larger estimation variance.

For the third model $\eta(x) = (x-10)^2/8$, it involves both noise and bias. From Figure 7, the estimation of $\phi(t) = 1/k$ is always above the true line, i.e., high bias, due to the convexity of the η. For interpolated-NN, their estimators are more rugged, but fluctuate along the true $\eta(x)$. For this case, it is difficult to claim which one is the best, but the trade-off phenomenon between bias and variance is still clear to see.

In conclusion, a fast increasing ϕ leads to smaller bias and larger variance, and non-interpolated weights such as k-NN leads to larger bias but smaller variance.

B Proof of Lemma 1

Proof. For simplicity, we write F_x as F, f_x as f, $v_p(x)$ as v_p and $\widehat{v}_p(x)$ as $\widehat{v}_p$ in this proof. Denote

$$A_2(a,b) = F(a) - \widehat{F}(a) - (F(b) - \widehat{F}(b)),$$

then for any $a, b \in \mathbb{R}^d$ with $|a-b|$ is larger than order $O(1/n)$, under Assumption (A.4), it follows that

$$\mathbb{E}A_2^{2m}(a,b) = O\bigg(\frac{(a-b)^m (f(a)+f(b))^m}{n^m}\bigg), \tag{9}$$

for any positive integer m.

To illustrate (9), for example, when $m = 4$, after expanding all the terms, the terms involving odd moments will be zero. The non-zero terms left are

$$\bigg[\mathbb{E}1_{\{\|X-x\| \in (a,b)\}} - (F(a) - F(b)))^2\bigg]^2 \tag{10}$$

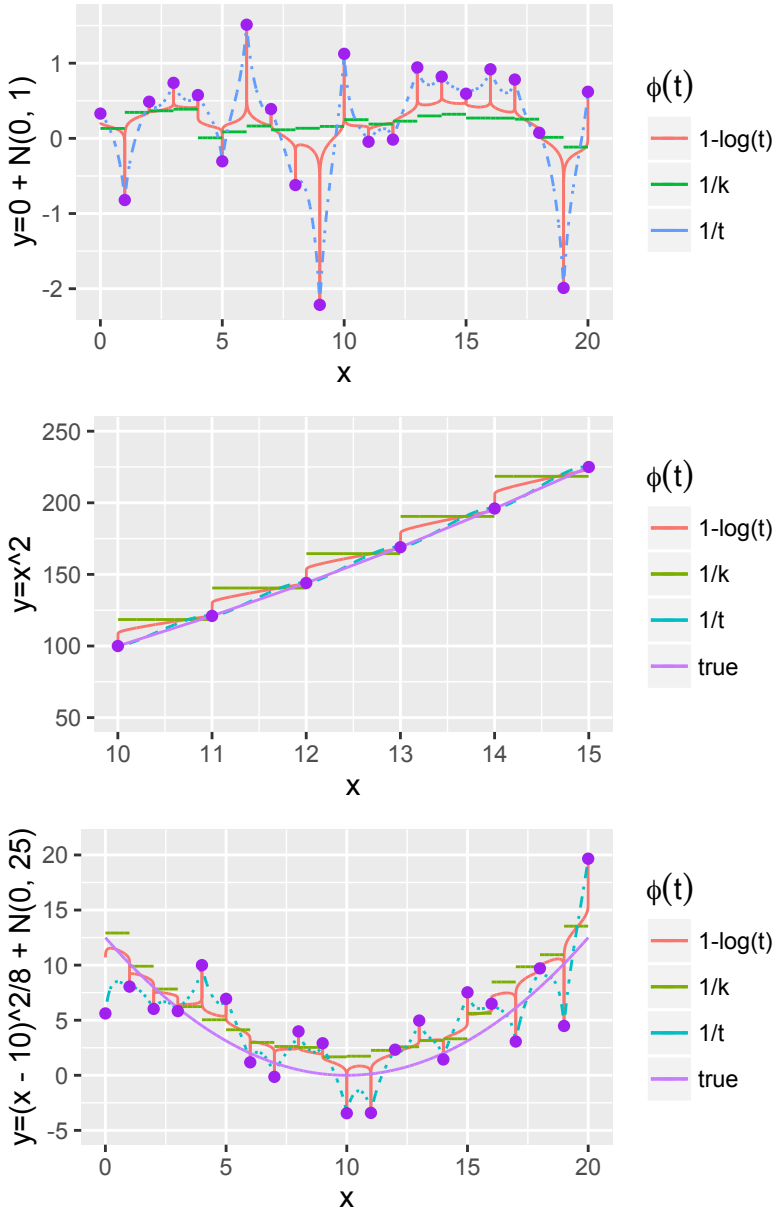

Figure 7: Three Trials on the Toy Simulation: Upper: $\eta \equiv 0$; Middle: $\eta(x) = x^2$; Lower: $\eta(x) = (x-10)^2/8$.

and

$$\mathbb{E}\bigg[1_{\{\|X-x\|\in(a,b)\}} - (F(a)-F(b))\bigg]^4. \tag{11}$$

For (10), there are $O(n^2)$ terms with this form, while for (11), there are only $O(n)$ terms. Since $a-b$ is larger than $O(1/n)$, the dominant part becomes (10). Therefore we can obtain (9) through further approximating $F(a)-F(b)$ as $[(a-b)(f(a)+f(b))]/2$.

Since $\widehat{F}(\widehat{v}_p) = F(v_p) = p$, we have

$$\widehat{F}(v_p) - \widehat{F}(\widehat{v}_p) = \widehat{F}(v_p) - F(v_p),$$

hence if $|\widehat{v}_p - v_p|$ and p is larger than $O(1/n)$,

$$\mathbb{E}\big[\widehat{F}(v_p) - \widehat{F}(\widehat{v}_p)\big]^{2m} = \mathbb{E}\big[\widehat{F}(v_p) - F(v_p)\big]^{2m} = O\bigg(\frac{p^m}{n^m}\bigg).$$

As a result, for $a, b > 0$, without loss of generality, assume $\widehat{v}_p > v_p$,

$$\begin{aligned}
& \mathbb{E}\big[F(v_p) - F(\widehat{v}_p)\big]^{2m} \\
\leq\ & P(\widehat{F}(\widehat{v}_p) - \widehat{F}(v_p) > b) + P(\widehat{F}(v_p+a) - \widehat{F}(v_p) < b) \\
& + \mathbb{E}\bigg[(F(v_p+a) - F(v_p))^{2m} 1_{\widehat{F}(\widehat{v}_p)\leq \widehat{F}(v_p+a)}\bigg] \\
\leq\ & P(\widehat{F}(\widehat{v}_p) - \widehat{F}(v_p) > b) + P(\widehat{F}(v_p+a) - \widehat{F}(v_p) < b) + \mathbb{E}\big[F(v_p+a) - F(v_p)\big]^{2m} \\
\leq\ & O\bigg(\frac{p^{m_2}}{b^{2m_2} n^{m_2}}\bigg) + O\big(a^{2m}(f(a+v_p)+f(v_p))^{2m}\big) + P(\widehat{F}(v_p+a) - \widehat{F}(v_p) < b) \\
=\ & O\bigg(\frac{p^{m_2}}{b^{2m_2} n^{m_2}}\bigg) + O\big(a^{2m}(f(a+v_p)+f(v_p))^{2m}\big) \\
& + P(F(a+v_p) - F(v_p) < b + A_2(v_p, v_p+a)) \\
\leq\ & O\bigg(\frac{p^{m_2}}{b^{2m_2} n^{m_2}}\bigg) + O\big(a^{2m}(f(a+v_p)+f(v_p))^{2m}\big) + O\bigg(\frac{b^{2m_1}}{a^{2m_1}(f(v_p)+f(a+v_p))^{2m_1}}\bigg) \\
& + O\bigg(\frac{1}{a^{m_1}(f(v_p)+f(a+v_p))^{m_1}} \frac{1}{n^{m_1}}\bigg).
\end{aligned}$$

When $n \to \infty$, taking $m_2 \gg m_1 \gg m$, we obtain

$$\mathbb{E}\big[F(v_p) - F(\widehat{v}_p)\big]^{2m} \leq O\bigg(\frac{p^m}{n^m}\bigg).$$

One can check that $a = o(v_p)$, hence $f(v_p) \approx f(v_p + a)$.

As a result, $\widehat{v}_p$ falls in the interval of $v_p \pm c$ with probability tending to 1 for some $c > 0$. We first assume that $c = o(v_p)$ hence $f(v_p + c) = O(f(v_p))$. Rewrite $f = f(v_p + c) f_(v_p)$ for simplicity, then

$$P(|\widehat{v}_p - v_p| > c) \leq O\bigg(\frac{p^m}{n^m} \frac{1}{c^{2m} f^{2m}}\bigg).$$

Therefore,

$$\mathbb{E}A_2(\widehat{v}_p, v_p)^2 \leq \mathbb{E}A_2^2(c+v_p, v_p) + O\bigg(\frac{p^m}{n^m} \frac{1}{c^{2m} f^{2m}}\bigg) = O\bigg(\frac{cf}{n}\bigg) + O\bigg(\frac{p^m}{n^m} \frac{1}{c^{2m} f^{2m}}\bigg).$$

The optimal rate becomes $O(p^{1/2}/n^{3/2})$ when $n \gg m \to \infty$. Note that $cf/n = O(p^{1/2}/n^{3/2})$ implies $c = O(p^{1/d} p^{-1/2}/n^{-1/2}) = o(p^{1/d}) = o(v_p)$. Since $p^{1/d-1/2}$ is of larger than $O(n^{-1/2})$, c is also larger than $O(1/n)$. □