Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate

https://bit.ly/38nizZW

accepted to NeurIPS2020 (poster)

Summary: Proposed multiscale *k*-NN improves the convergence rate of *k*-NN.

Akifumi Okuno[1,3] and Hidetoshi Shimodaira[2,3]

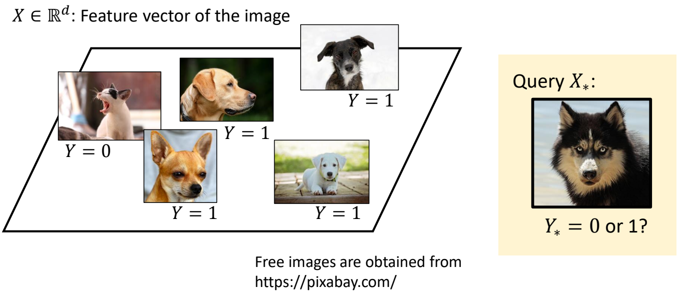
[1]Inst. of Stat. Math. [2]Kyoto Univ. [3]RIKEN AIP

Table of contents

Preliminaries

Classification problem

- Let $X \in \mathbb{R}^d$, $Y \in \{0,1\}$ be random variables, where $(X, Y) \sim \mathbb{Q}$.
- *Observations* $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1}^n$ are independent copies of (X, Y).

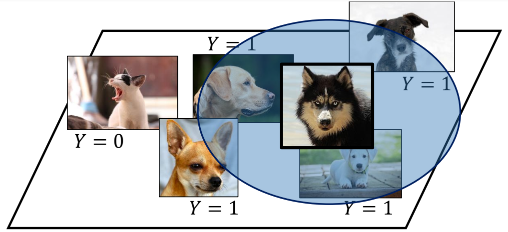


- We aim at obtaining a *classifier* $\hat{g}_n : \mathbb{R}^d \to \{0,1\}$ that minimizes

$$\mathbb{P}_{(X_*, Y_*) \sim \mathbb{Q}}(Y_* \neq \hat{g}_n(X_*)),$$

where X_* is a *query*, and $\hat{g}_n$ is trained with the observations $\mathcal{D}_n$.

k-nearest neighbour (k-NN) classifier



- Rearrange the index such that

$$\|X_* - X_{(1)}\|_2 \le \|X_* - X_{(2)}\|_2 \le \cdots \le \|X_* - X_{(n)}\|_2.$$

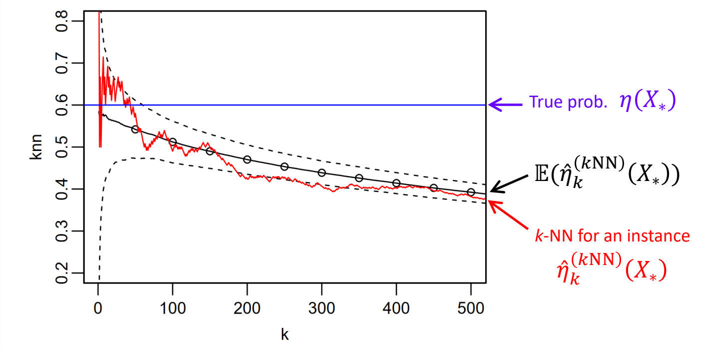
- *k-NN estimator* is defined by the ratio

$$\hat{\eta}_k^{(k\mathrm{NN})}(X_*) := k^{-1} \sum_{i=1}^{k} Y_{(i)}.$$

- Hereinafter, we only consider plug-in classifier $\hat{g}(X_*) = \mathbb{1}(\hat{\eta}(X_*) \ge 1/2)$ defined for estimators $\hat{\eta}$.

Bias-variance tradeoff

- small k: ☺ small bias ☹ large variance
- large k: ☺ small variance ☹ large bias



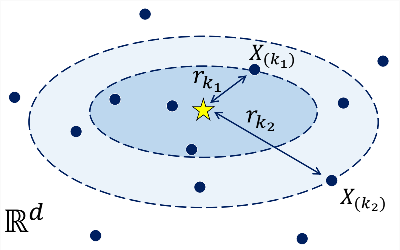
- **Conventional**: choose the best k value by considering the tradeoff.
- **Ours**: reduces the asymptotic bias!

Proposal: multiscale k-NN

How to reduce the bias?: An overview

Consider a radius $r_k := \|X_* - X_{(k)}\|_2$, a ball $B(X; r)$, and $\eta^{(\infty)}(B) := \mathbb{E}(Y \mid X \in B)$, where Chaudhuri and Dasgupta (2014) proves that

$$\hat{\eta}_k^{(k\text{NN})}(X_*) \approx \eta^{(\infty)}(B(X_*, r_k)) \quad (k = k_n \to \infty, n \to \infty \text{ and } k/n \to 0).$$



bias is due to $r > 0 \Rightarrow$ (imaginary) $r = 0$ is preferred:

$$\eta^{(\infty)}(B(X_*; r)) \to \eta(X_*) = \mathbb{E}[Y_* \mid X_*], \quad (r \to 0;\ \text{Federer (1967)})$$

To obtain **(imaginary)** 0-**NN estimator**, we extrapolate k-NN estimators

$$\hat{\eta}_{k_1}^{(k\text{NN})}(X_*), \hat{\eta}_{k_2}^{(k\text{NN})}(X_*), \dots, \hat{\eta}_{k_V}^{(k\text{NN})}(X_*)$$

to $r = 0$ via the radii $r_{k_1}, r_{k_2}, \dots, r_{k_V}$.

Multiscale k-NN

Consider a set $\mathcal{F}$ of regression functions $f : \mathbb{R} \to \mathbb{R}$ (e.g., $f(r) = \beta_0 + \beta_1 r$). By choosing $V \in \mathbb{N}$ and $1 \leq k_1 < k_2 < \cdots < k_V \leq n$, we conduct a regression

$$\hat{f} := \mathop{\arg\min}_{f \in \mathcal{F}} \sum_{v=1}^{V} \left(\hat{\eta}_{k_v}^{(kNN)} - f(r_{k_v}) \right)^2, \quad (r_k := \|X_* - X_{(k)}\|_2).$$

Definition (Multiscale k-NN)

*We define a **multiscale k-NN** (MS-k-NN) estimator*

$$\hat{\eta}_{\boldsymbol{k}}^{(MSkNN)}(X_*) := \hat{f}(0)$$

where $\boldsymbol{k} = (k_1, k_2, \ldots, k_V)$.

MS-k-NN formally extrapolates k-NN estimators to $r = 0$.

Comparison with existing estimators

Comparison with k-NN estimator

Roughly speaking, bias is reduced for $\beta > 2$:

$$k\text{-NN:}\quad |\hat{\eta}_k^{(k\text{NN})}(X_*) - \eta(X_*)| \approx O(r_k^2),$$

$$\text{MS-}k\text{-NN:}\quad |\hat{\eta}_{\boldsymbol{k}}^{(\text{MS}k\text{NN})}(X_*) - \eta(X_*)| \approx O(r_k^{\beta}).$$

Variances are in the same order; overall, the convergence rate is reduced.

Comparison with local polynomial (LP) estimator

LP and MS-k-NN attain the **same optimal convergence rate**. However, MS-k-NN requires **much less terms** than LP, to obtain the same rate.

LP: $1 + d + d^2 + \cdots + d^C$ coefficients, to estimate Taylor polynomial of $\eta(X)$ (and extrapolate to X_*),

MS-k-NN: $1 + C$ coefficients to be estimated.

Furthermore, MS-k-NN is also expected to inherit the favorable properties of k-NN.

Theories: convergence rate analysis

Convergence rate of the excess risk

Given a classifier $g : \mathbb{R}^d \to \{0,1\}$, we define a misclassification error rate $L(g) := \mathbb{P}_{(X_*,Y_*)\sim\mathbb{Q}}(Y_* \neq g(X_*))$ and excess risk

$$\mathcal{E}(g) := L(g) - \inf_{g:\mathbb{R}^d\to\{0,1\}} L(g).$$

Convergence rate:

the order of $\mathcal{E}(\hat{g}_n)$ w.r.t. n.

In order to elucidate the convergence rate, we employ

- α-margin condition,
- β-Hölder condition, and
- γ-neighbour average smoothness condition,

by referring to existing studies (see, e.g., Audibert and Tsybakov (2007), Samworth (2012) and Chaudhuri and Dasgupta (2014)).

Definition (α-margin condition)

If $\exists L_\alpha > 0, \tilde{t} > 0, \alpha \geq 0$ such that

$$\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq L_\alpha t^\alpha \quad (\forall t \in (0, \tilde{t}], X \in \mathcal{X}),$$

*η is said to be satisfying α-**margin condition**.*

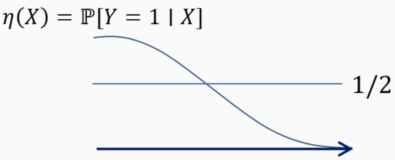


Figure: Small α

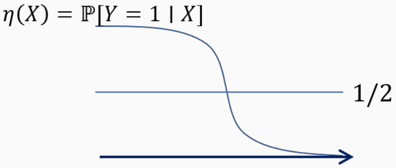


Figure: Large α

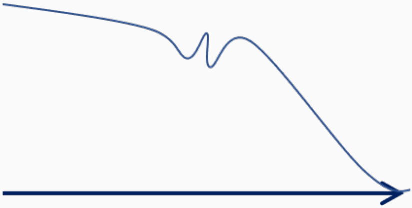
α is Large $\Rightarrow$ only a few covariates are near boundary $\Rightarrow$ classification is easy.
$\Rightarrow$ fast convergence

Definition (β-Hölder condition)

Let $\mathcal{T}_{q,X_*}[\eta]$ be the Taylor expansion of η of degree $q \in \mathbb{N}_0$. If $\exists L_\beta > 0$ such that

$$|\eta(X) - \mathcal{T}_{\lfloor\beta\rfloor,X_*}[\eta](X)| \leq L_\beta \|X - X_*\|^\beta \quad (\forall X, X_* \in \mathcal{X}),$$

*η is said to be satisfying β-**Hölder condition**.*



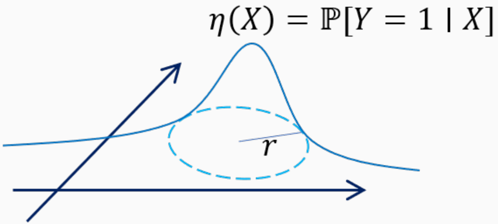
β is large $\Rightarrow$ $\eta(X)$ is smooth $\Rightarrow$ estimation of η is easy
$\Rightarrow$ fast convergence

Definition (γ-neighbour average smoothness condition)

Let $\eta^{(\infty)}(B) := \mathbb{E}[Y \mid X \in B]$*. If* $\exists L_\gamma, \gamma > 0$ *such that*

$$|\eta^{(\infty)}(B(X;r)) - \eta(X)| \le L_\gamma r^\gamma \quad (\forall r > 0, X \in \mathcal{S}(\mu)),$$

η *is said to be satisfying* γ***-neighbour average smoothness condition****.*



γ is large $\Rightarrow$ k-NN approximation $\eta^{(\infty)}(B(X_*;r)) = \mathbb{E}[Y \mid X \in B(X_*;r)]$ converges to $\mathbb{E}[Y \mid X_*]$ quickly
$\Rightarrow$ fast convergence

Convergence rate of k-NN

Definition (Strong density assumption (SDA) on pdf μ of X)

If $\exists \mu_{\min}, \mu_{\max} \in (0, \infty)$ such that $\mu(X) \in [\mu_{\min}, \mu_{\max}]$ for all $X \in \mathcal{X}$, μ is said to be satisfying ***strong density assumption*** *(SDA).*

Theorem (Chaudhuri and Dasgupta (2014) Theorem 4)

Let $\mathcal{X}$ be a compact set, and assuming that (i) η satisfies α- and γ- conditions, and (ii) μ satisfies SDA, it holds with $k_* \asymp n^{2\gamma/(2\gamma+d)}$ that

$$\mathcal{E}(\hat{g}_{k_*}^{(k\mathrm{NN})}) = O(n^{-(1+\alpha)\gamma/(2\gamma+d)}),$$

for (unweighted) k-NN plug-in classifier $\hat{g}_{k_*}^{(k\mathrm{NN})}$.

- A natural question: $\gamma = \beta$ if β-Hölder condition is employed instead of γ-?

Answer is **No**.

Theorem (Okuno and Shimodaira (2020) Theorem 1)

Let $\eta(X) := \mathbb{E}(Y \mid X)$ and let μ be the p.d.f. of X. Assuming that

(1) both of μ and $\eta\mu$ are $\beta(>0)$-Hölder,

(2) support of μ is compact,

(3) $k = k_n \to \infty, k/n \to 0, n \to \infty$.

Then, for some $b_2^*, \dots, b_{\lfloor \beta/2 \rfloor}^* \in \mathbb{R}$, it holds that

$$\eta^{(\infty)}(B(X_*, r_k)) = \eta(X_*) + \underbrace{b_1^* r_k^2 + b_2^* r_k^4 + \cdots + b_{\lfloor \beta/2 \rfloor}^* r_k^{2\lfloor \beta/2 \rfloor} + O(r_k^\beta)}_{\text{bias}},$$

and $b_1^* = \frac{1}{2d+4} \frac{1}{\mu(X_*)} \{\Delta[\eta(X_*)\mu(X_*)] - \eta(X_*)\Delta\mu(X_*)\}$ with the Laplacian operator Δ.

If β-Hölder condition is assumed instead of γ-, we have

$$\gamma = \min\{\beta, 2\},$$

indicating that $\mathcal{E}(\hat{g}_{k_*}^{(k\mathrm{NN})}) = O(n^{-2(1+\alpha)/(4+d)})$ even for sufficiently smooth η.

Optimal rate for plug-in classifiers

Audibert and Tsybakov (2007) Theorem 3.5 proves the optimal rate for plug-in classifiers: considering $\beta(>0)$-Hölder function η, there exists $L > 0$ such that

$$\inf_{g:\text{plug-in classifier}} \sup_{\eta,\mu} \mathcal{E}(g) \geq L \cdot n^{-(1+\alpha)\beta/(2\beta+d)}.$$

Table: Convergence rates for $\alpha = 1, \beta = 2u$ ($u \in \mathbb{N}$)

k-NN	$O(n^{-4/(4+d)})$	Chaudhuri and Dasgupta (2014)
Local linear	$O(n^{-4/(4+d)})$	Hall and Kang (2005)
Local polynomial	$O(n^{-2\beta/(2\beta+d)})$	Audibert and Tsybakov (2007)
Multiscale *k*-NN	$O(n^{-2\beta/(2\beta+d)})$	Okuno and Shimodaira (2020)

Another implication of Theorem 1

Okuno and Shimodaira (2020) Theorem 1:

$$\underbrace{\eta^{(\infty)}(B(X_*, r_k))}_{k\text{NN estimator}} = \eta(X_*) + \underbrace{b_1^* r_k^2 + b_2^* r_k^4 + \cdots + b_{\lfloor \beta/2 \rfloor}^* r_k^{2\lfloor \beta/2 \rfloor} + O(r_k^\beta)}_{\text{bias}}$$

leads to a regression function

$$f_C(r; \boldsymbol{b}) := b_0 + b_1 r^2 + b_2 r^4 + \cdots + b_C r^{2C} \quad (C = \lfloor \beta/2 \rfloor).$$

The function f_C approximates the bias term, and extrapolation to $r = 0$ yields

$$f_C(0; \hat{\boldsymbol{b}}) = \hat{b}_0 \approx \eta(X_*).$$

Therefore, we may employ a set of even-degree polynomials

$$\mathcal{F}_C := \{b_0 + b_1 r_1^2 + b_2 r_2^4 + \cdots + b_C r^{2C} \mid b_0, b_1, \ldots, b_C \in \mathbb{R}\}.$$

MS-k-NN attains the optimal rate!

With user-specified $\ell_1 = 1 < \ell_2 < \cdots < \ell_V < \infty$, we consider

(C1) $k_1 \asymp n^{2\beta/(2\beta+d)}$,

(C2) $k_v := \min\{k \in [n] \mid \|X_{(k)} - X_*\|_2 \geq \ell_v r_{k_1}\}$ for $v = 2, 3, \ldots, V$,

(C3) $\exists L_z > 0$ such that $\|\frac{(I-\mathcal{P}_{\boldsymbol{R}})\mathbf{1}}{\mathbf{1}^\top(I-\mathcal{P}_{\boldsymbol{R}})\mathbf{1}}\|_\infty \leq L_z$ for $\boldsymbol{R} = (\ell_i^{2j})_{ij}, \mathcal{P}_{\boldsymbol{R}} = \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1}\boldsymbol{R}^\top$.

Theorem (Okuno and Shimodaira (2020) Theorem 2)

Assuming that (i) $\mu, \eta\mu$ are β-Hölder[1], (ii) μ satisfies SDA, (iii) $C := \lfloor \beta/2 \rfloor \leq V - 1$, and (iv) (C-1)–(C-3) are satisfied. Then, MS-k-NN plug-in classifier attains the optimal rate

$$\mathcal{E}(\hat{g}_{\boldsymbol{k}_*}^{(\mathrm{MS}k\mathrm{NN})}) = O(n^{-(1+\alpha)\beta/(2\beta+d)}).$$

[1] For k-NN, only η is assumed to be β-Hölder (Chaudhuri and Dasgupta, 2014)

Weighted k-NN

Consider a **weighted** k-NN estimator

$$\hat{\eta}^{(k\mathrm{NN})}_{k,\boldsymbol{w}}(X_*) = \sum_{i=1}^{k} w_i Y_{(i)}$$

with weights $\sum_{i=1}^{k} w_i = 1$. Then, MS-k-NN is equivalent to the weighted k-NN equipped with $k = k_V$ and **real-valued weights**

$$w_i := \sum_{v: i \le k_v} \frac{z_v}{k_v} \in \mathbb{R} \ (\forall i \in [k_V]), \quad \boldsymbol{z} = (z_1, z_2, \ldots, z_V) := \frac{(I - \mathcal{P}_{\boldsymbol{R}})\mathbf{1}}{\mathbf{1}^\top (I - \mathcal{P}_{\boldsymbol{R}})\mathbf{1}} \in \mathbb{R}^V.$$

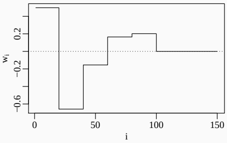


Figure: $V = 5$

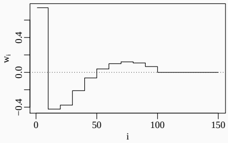


Figure: $V = 10$

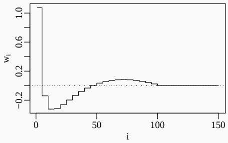


Figure: $V = 20$

- **Negative weights** are essential for eradicating the bias.

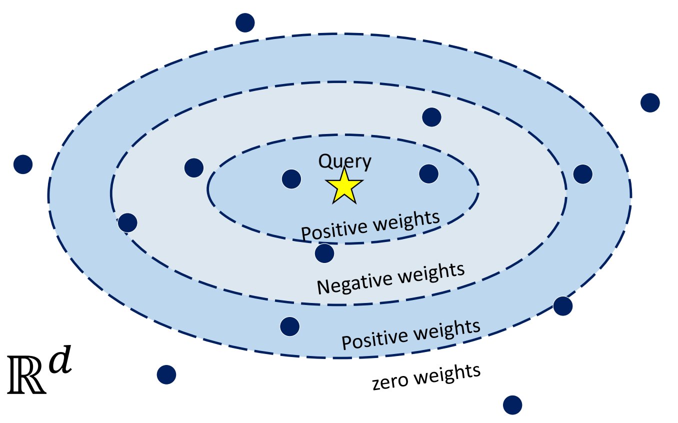


Figure: (Implicit) weights obtained in multiscale k-NN

Only one existing study that considers k-NN with real-valued weights is Samworth (2012); it proves for $\alpha = 1, k_* \asymp n^{2\beta/(2\beta+d)}$ that

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \mathcal{E}(\hat{g}^{(\mathrm{kNN})}_{k_*,\boldsymbol{w}}) = O(n^{-(1+\alpha)\beta/(2\beta+d)}) \text{ for a conditioned set } \mathcal{W} \subseteq \mathbb{R}^{k_*}.$$

Samworth (2012) shows equations of the optimal weights by minimizing Taylor series of the excess risk; solutions are obtained only for $\beta = 2, 4$. For $\beta = 4$,

$$w_i = (a_0 + a_i\delta_i^{(1)} + \cdots + a_u\delta_i^{(u)})/k_*$$

with $\delta_i^{(\ell)} := i^{1+2\ell/d} - (i-1)^{1+2\ell/d}$ $(\forall \ell \in [u])$, $a_0 \in \mathbb{R}$, $a_1 := \frac{1}{k_*^{2/d}}\{\frac{(d+4)^2}{4} - \frac{2(d+4)}{d+2}a_0\}$, and $a_2 := \frac{1-a_0-k_*^{2/d}a_1}{k_*^{4/d}}$.

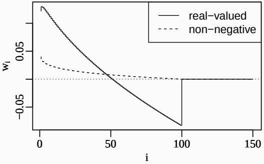


Figure: Optimal real-valued weights in Samworth (2012)

- cf. Samworth (2012) also proves the rate $O(n^{-4/(4+d)})$ for non-negative weights, which is the same as unweighted k-NN.

Numerical experiments

Problem setup

We perform

(1) unweighted k-NN ($w_i = 1/k$)
(2) weighted k-NN with *non-negative* weights $w_i \geq 0$
(3) weighted k-NN with *real-valued* weights $w_i \in \mathbb{R}$
(4) (**Proposal**) MS-k-NN extrapolated via $r(k)$
(5) (**Modificaiton**) MS-k-NN extrapolated via $\log k$

on 13 datasets obtained from UCI ML Repository (Dua and Graff, 2017).

- Divided into: 70% for training, 30% for test.
- Sample mean and standard deviation of the prediction accuracy on 10 times experiments are computed.
- Regression in MS-k-NN is ridge regularized with $\lambda = 10^{-4}$.
- $V = 5, k := n_{\text{train}}^{4/(4+d)}, k_1 = k/V, k_2 = 2k/V, ..., k_V = k$.

Prediction accuracy

- n: number of observations
- d: dimension of X
- m: number of categories

Table: Best scores are **bolded**, and second best scores are underlined.

Dataset	n	d	m	k-NN			**MS-k-NN**	
				$w_i = 1/k$	$w_i \geq 0$	$w_i \in \mathbb{R}$	via $r(k)$	via $\log k$
Iris	150	4	3	0.83 ± 0.04	0.92 ± 0.05	0.92 ± 0.04	$\underline{0.93} \pm 0.04$	$\mathbf{0.96} \pm 0.04$
Glass iden.	213	9	6	0.58 ± 0.06	$\underline{0.64} \pm 0.06$	$\mathbf{0.67} \pm 0.05$	$\underline{0.64} \pm 0.05$	$\underline{0.64} \pm 0.05$
Ecoli	335	7	8	0.80 ± 0.03	$\mathbf{0.85} \pm 0.03$	$\underline{0.84} \pm 0.02$	$\mathbf{0.85} \pm 0.02$	$\underline{0.84} \pm 0.02$
Diabetes	768	8	2	$\mathbf{0.75} \pm 0.03$	$\underline{0.74} \pm 0.03$	0.70 ± 0.04	$\mathbf{0.75} \pm 0.03$	0.71 ± 0.03
Biodeg.	1054	41	2	$\underline{0.84} \pm 0.02$	$\mathbf{0.86} \pm 0.03$	0.79 ± 0.02	$\mathbf{0.86} \pm 0.02$	0.80 ± 0.02
Banknote	1371	4	2	0.95 ± 0.01	$\underline{0.98} \pm 0.01$	0.97 ± 0.01	$\underline{0.98} \pm 0.01$	$\mathbf{0.99} \pm 0.00$
Yeast	1484	8	10	$\underline{0.57} \pm 0.02$	$\mathbf{0.58} \pm 0.02$	0.54 ± 0.03	$\mathbf{0.58} \pm 0.02$	0.54 ± 0.02
Wire. local.	2000	7	4	$\underline{0.97} \pm 0.00$	$\mathbf{0.98} \pm 0.00$	$\mathbf{0.98} \pm 0.01$	$\mathbf{0.98} \pm 0.00$	$\mathbf{0.98} \pm 0.01$
Spambase	4600	57	2	$\underline{0.90} \pm 0.01$	$\mathbf{0.91} \pm 0.00$	0.86 ± 0.01	$\mathbf{0.91} \pm 0.00$	0.87 ± 0.01
Robot navi.	5455	24	4	0.81 ± 0.01	$\mathbf{0.86} \pm 0.01$	0.81 ± 0.01	$\underline{0.84} \pm 0.01$	$\underline{0.84} \pm 0.01$
Page blocks	5473	10	5	$\underline{0.95} \pm 0.01$	$\underline{0.95} \pm 0.01$	$\mathbf{0.96} \pm 0.01$	$\mathbf{0.96} \pm 0.01$	$\mathbf{0.96} \pm 0.01$
MAGIC	19020	10	2	0.82 ± 0.00	0.82 ± 0.00	$\mathbf{0.84} \pm 0.01$	$\underline{0.83} \pm 0.00$	$\underline{0.83} \pm 0.00$
Avila	20867	10	12	0.63 ± 0.01	0.68 ± 0.01	$\mathbf{0.70} \pm 0.01$	$\underline{0.69} \pm 0.00$	$\mathbf{0.70} \pm 0.01$

Some remarks

Okuno and Shimodaira (2020) Theorem 1 proves that

$$\eta^{(\infty)}(B(X_*; r_k) = \eta(X_*) + \sum_{c=1}^{\lfloor \beta/2 \rfloor} r_k^{2\lfloor \beta/2 \rfloor} + O(r_k^{\beta}).$$

for **small** $r_k \approx 0$ (as $k/n \to 0, n \to \infty$).

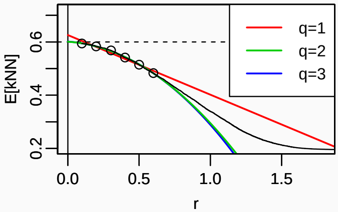


Figure: $\delta = 0.1$

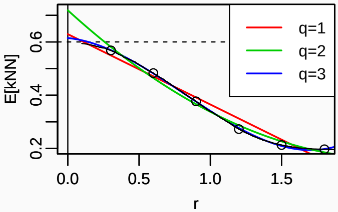


Figure: $\delta = 0.3$

Figure: Monte-Carlo expectation of k-NN estimators (black line), and the polynomials of degrees $q = 1, 2, 3$ trained on $r = \delta, 2\delta, \ldots, 6\delta$.

- Even degree polynomials $b_0 + b_1 r_1^2 + \cdots + b_C r^{2C} : \mathbb{R} \to \mathbb{R}$ can be replaced with

$$\sigma\left(b_0 + b_1 r_1^2 + \cdots + b_C r^{2C}\right) \; : \; \mathbb{R} \to [0, 1]$$

using the sigmoid function $\sigma(z) = (1 + \exp(-z))^{-1}$, to attain the optimal rate. (These two functions are essentially equivalent for small $r \approx 0$.)

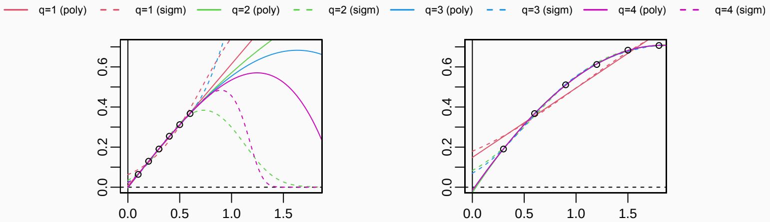


Figure: $\delta = 0.1$ Figure: $\delta = 0.3$

Figure: Sigmoid-based functions (dot lines).

Theorem 1 revisited

$$\eta^{(\infty)}(B(X_*;r)) = \mathbb{E}[Y \mid X \in B(X_*;r)] = \frac{\int_{B(X_*;r)} \eta(X)\mu(X)\mathrm{d}X}{\int_{B(X_*;r)} \mu(X)\mathrm{d}X}$$

- To apply Taylor-expansion, $\mu, \eta\mu$ are assumed to be β-Hölder in Theorem 1:

$$\eta^{(\infty)}(B(X_*;r)) = \sum_{c=0}^{C} b_c^* r^{2c} + O(r^\beta).$$

- If $\mu, \eta\mu$ are polynomial, we have a non-asymptotic expansion:

$$\eta^{(\infty)}(B(X_*;r)) = \mathbb{1}(C_1 - C_2 \geq 0) \sum_{c=0}^{C_1 - C_2} b_c^* r^{2c} + \frac{\sum_{c=0}^{C_2-1} \gamma_c^{(1)} r^{2c}}{\sum_{c=0}^{C_2} \gamma_c^{(2)} r^{2c}}$$

for some $\{b_c^*\}, \{\gamma_c^{(1)}\}, \{\gamma_c^{(2)}\} \subset \mathbb{R}$.

k-NN estimators $\hat{\eta}^{(kNN)}_{k_1}, \hat{\eta}^{(kNN)}_{k_2}, \ldots, \hat{\eta}^{(kNN)}_{k_V}$ are dependent.

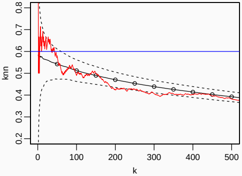


Figure: k-NN

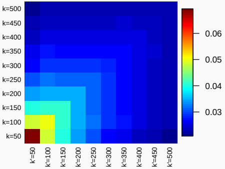


Figure: covariance$^{1/2}$

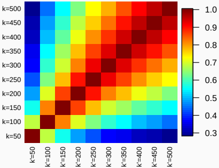


Figure: Correlation

Figure: Dependence of k-NN estimators computed via Monte-Carlo simulation.

- Dependence can be considered in the regression.

Choosing parameters

- Cross-validation is conducted for choosing the parameters $k_1, k_2, \ldots, k_V$.
- Instead of choosing $1 \leq k_1 < k_2 < \cdots < k_V \leq n$, we may employ $k_1 = 1, k_2 = 2, \ldots, k_{V'} = V'$ ($V \ll V'$; for avoiding parameter selection): empirically better performance in some cases.

Conclusion

Conclusion

- To obtain **(imaginary) 0-NN estimator**, k-NN estimators $\hat{\eta}_{k_1}, \hat{\eta}_{k_2}, \ldots, \hat{\eta}_{k_V}$ are extrapolated to $r = 0$ via radius $r_k := \|X_{(k)} - X_*\|_2$.
- Obtained multiscale k-NN (MS-k-NN) estimator reduces the bias of k-NN, and **it attains the optimal rate**.
- MS-k-NN is equivalent to weighted k-NN with some real-valued weights.
- Weights are automatically determined via regression (in MS-k-NN); they are different from Samworth (2012), which solves entangled equations.

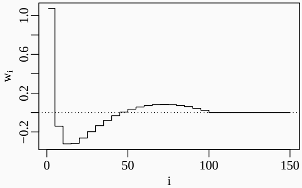


Figure: Ours ($V = 20$)

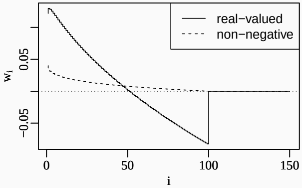


Figure: Samworth (2012)

Acknowledgement: We would like to thank Bourbon Cao (Kyoto U.) and Takuma Tanaka (Kyoto U.) for helpful discussions.

References I

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633.

Chaudhuri, K. and Dasgupta, S. (2014). Rates of convergence for nearest neighbor classification. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3437–3445. Curran Associates, Inc.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer, New York.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository.

Federer, H. (1967). *Geometric measure theory*. Springer.

Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.*, 33(1):284–306.

Okuno, A. and Shimodaira, H. (2020). Extrapolation towards imaginary 0-nearest neighbour and its improved convergence rate. *arXiv preprint arXiv:2002.03054*.

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