Nonparametric Analysis: Nearest Neighbors and Friends

Samory Kpotufe
Statistics, Columbia University
Nonparametric Analysis:

Infinite capacity/number of parameters $\not\Rightarrow$ no Generalization

Which aspects of a procedure/data, $\Rightarrow$ fast/slow Generalization

...
Nonparametric Analysis:

Infinite capacity/number of parameters $\Rightarrow$ no Generalization

Which aspects of a procedure/data, $\Rightarrow$ fast/slow Generalization

...
*k*-Nearest Neighbor Approach:
Use the *k* closest datapoints to *x* to infer something about *x*.

Ubiquitous in ML (implicit at times):

**Traditional ML:** Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

**Modern ML:** Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:
Which metric? Which values of *k*? Implementation and Tradeoffs?

A lot of recent insights towards these questions ...
$k$-Nearest Neighbor Approach:
Use the $k$ closest datapoints to $x$ to infer something about $x$.

Ubiquitous in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ... 
Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:
Which metric? Which values of $k$? Implementation and Tradeoffs?
A lot of recent insights towards these questions ...
$k$-Nearest Neighbor Approach:
Use the $k$ closest datapoints to $x$ to infer something about $x$.

Ubiquitous in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...  
Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:
Which metric? Which values of $k$? Implementation and Tradeoffs?
A lot of recent insights towards these questions ...
\textbf{\textit{k}-Nearest Neighbor Approach:}

Use the \textit{k closest} datapoints to \textit{x} to \textit{infer} something about \textit{x}.

\textbf{Ubiquitous in ML (implicit at times):}

\textbf{Traditional ML:} Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...  
\textbf{Modern ML:} Matrix Completion, Inference on Graphs, Time Series Prediction ...

\textbf{Of Practical Interest:}

Which metric? Which values of \textit{k}? Implementation and Tradeoffs? A lot of recent insights towards these questions ...
\textit{k-Nearest Neighbor Approach:}

Use the $k$ closest datapoints to $x$ to infer something about $x$.

\textbf{Ubiquitous in ML (implicit at times):}

\textbf{Traditional ML:} Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

\textbf{Modern ML:} Matrix Completion, Inference on Graphs, Time Series Prediction ...

\textbf{Of Practical Interest:}

Which metric? Which values of $k$? Implementation and Tradeoffs?

A lot of recent insights towards these questions ...
$k$-Nearest Neighbor Approach:
Use the $k$ closest datapoints to $x$ to infer something about $x$.

Ubiquitous in ML (implicit at times):

**Traditional ML:** Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...  
**Modern ML:** Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:
Which metric? Which values of $k$? Implementation and Tradeoffs?

A lot of recent insights towards these questions ...
**k-Nearest Neighbor Approach:**

Use the \( k \) closest datapoints to \( x \) to infer something about \( x \).

**Ubiquitous in ML (implicit at times):**

**Traditional ML:** Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

**Modern ML:** Matrix Completion, Inference on Graphs, Time Series Prediction ...

**Of Practical Interest:**

Which metric? Which values of \( k \)? Implementation and Tradeoffs?

A lot of recent insights towards these questions ...
Basic Intuition:

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

\[
\begin{array}{cccccc}
1 & 1 & 5 & 4 & 3 \\
7 & 5 & 3 & 5 & 3 \\
5 & 9 & 0 & 6 \\
3 & 5 & 2 & 0 & 0 \\
\end{array}
\]

... $y \leftarrow 5$

Prediction: aggregate $Y$ values in Neighborhood$(x)$

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition:

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, von Luxburg, Hein, Chaudhuri, Dasgupta ...
Basic Intuition:

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

Prediction: aggregate $Y$ values in Neighborhood($x$)

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, von Luxburg, Hein, Chaudhuri, Dasgupta ...
Basic Intuition:

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

Prediction: aggregate $Y$ values in Neighborhood$(x)$

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, von Luxburg, Hein, Chaudhuri, Dasgupta ...
Basic Intuition:

Closest neighbors of \( x \) should be mostly of similar type \( y = y(x) \) ...

\[
x \equiv 5
\]

\[
\cdots \quad y \leftarrow 5
\]

Prediction: aggregate \( Y \) values in Neighborhood(\( x \))

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition:

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Gy"orfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, von Luxburg, Hein, Chaudhuri, Dasgupta ...
Basic Intuition:

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

Prediction: aggregate $Y$ values in $\text{Neighborhood}(x)$

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition:

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, von Luxburg, Hein, Chaudhuri, Dasgupta ...
Key questions:

1 Statistical issues: how well can NN perform?
   - When is 1-NN enough?
   - For $k$-NN, what should $k$ be?
   - Is there always a curse of dimension?

2 Algorithmic issues: how efficient can NN be?
   - Which data structure to use?
   - Can we parallelize NN?
   - What do we tradeoff?
Key questions:

1 Statistical issues: how well can NN perform?
   • When is 1-NN enough?
   • For $k$-NN, what should $k$ be?
   • Is there always a curse of dimension?

2 Algorithmic issues: how efficient can NN be?
   • Which data structure to use?
   • Can we parallelize NN?
   • What do we tradeoff?
Key questions:

1 **Statistical issues**: how well can NN perform?
   - When is 1-NN enough?
   - For \( k \)-NN, what should \( k \) be?
   - Is there always a curse of dimension?

2 **Algorithmic issues**: how efficient can NN be?
   - Which data structure to use?
   - Can we parallelize NN?
   - What do we tradeoff?
Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

... 

Representation $\equiv$ choice of metric or dissimilarity $\rho(x, x')$

Properties of $\rho$ influence Statistical and Algorithmic aspects
Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

Representation $\equiv$ choice of metric or dissimilarity $\rho(x, x')$

Properties of $\rho$ influence Statistical and Algorithmic aspects...
Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

Representation ≡ choice of metric or dissimilarity $\rho(x, x')$
Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

... 

\[ \text{Representation} \equiv \text{choice of metric or dissimilarity} \quad \rho(x, x') \]

Properties of \( \rho \) influence Statistical and Algorithmic aspects 

...
Tutorial Outline:

- **PART I:** Basic Statistical Insights
- **PART II:** Best Practice and Tradeoffs
Tutorial Outline:

- **PART I:** Basic Statistical Insights
- **PART II:** Best Practice and Tradeoffs
PART I: Basic Statistical Insights

• Universality

• Behavior of $k$-NN Distances

• From Regression to Classification

• Classification is easier than regression

• Multiclass and Mixed Costs
$k$-NN as a universal approach:
it can fit anything, provided $k$ grows (but not too fast) with sample size!

Let's make this precise in the context of regression ...

For simplicity, assume $P_X$ is continuous on $\mathbb{R}^d$ ... (i.e. no ties)
$k$-NN as a universal approach:

it can fit anything, provided $k$ grows (but not too fast) with sample size!

Let’s make this precise in the context of regression ...

For simplicity, assume $P_X$ is continuous on $\mathbb{R}^d$ ... (i.e. no ties)
$k$-NN as a universal approach:
it can fit anything, provided $k$ grows (but not too fast) with sample size!

Let’s make this precise in the context of regression ...

For simplicity, assume $P_X$ is continuous on $\mathbb{R}^d$ ... (i.e. no ties)
**k-NN Regression**

**i.i.d. Data:** \( \{(X_i, Y_i)\}_{i=1}^{n} \),

\( Y = f(X) + \text{noise} \)

Learn: \( f_k(x) = \text{avg} \ (Y_i) \text{ of } k\text{-NN}(x) \).

**k-NN is universally consistent:**

Suppose \( \frac{k}{n} \to 0 \) and \( k \to \infty \), then \( \mathbb{E} |f_k(X) - f(X)| \xrightarrow{n \to \infty} 0 \)

Any function \( f, \mathbb{E}f^2 < \infty \), no matter how complex.
**k-NN Regression**

**i.i.d. Data:** \( \{(X_i, Y_i)\}_{i=1}^n \),

\( Y = f(X) + \text{noise} \)

**Learn:** \( f_k(x) = \text{avg} \ (Y_i) \text{ of } k-\text{NN}(x) \).

**k-NN is universally consistent:**

Suppose \( \frac{k}{n} \to 0 \) and \( k \to \infty \), then \( \mathbb{E} |f_k(X) - f(X)| \to 0 \)

Any function \( f, \mathbb{E}f^2 < \infty \), no matter how complex.
**$k$-NN Regression**

**i.i.d. Data:** \( \{(X_i, Y_i)\}_{i=1}^{n} \),
\[ Y = f(X) + \text{noise} \]

**Learn:** \( f_k(x) = \text{avg} (Y_i) \) of \( k\)-NN\( (x) \).

**$k$-NN is universally consistent:**

Suppose \( \frac{k}{n} \to 0 \) and \( k \to \infty \), then
\[ \mathbb{E} |f_k(X) - f(X)| \xrightarrow{n \to \infty} 0 \]

Any function \( f, \mathbb{E} f^2 < \infty \), no matter how complex.
$k$-NN Regression

i.i.d. Data: $\{(X_i, Y_i)\}_{i=1}^{n}$, $Y = f(X) + \text{noise}$

Learn: $f_k(x) = \text{avg}(Y_i)$ of $k$-NN$(x)$.

$k$-NN is universally consistent:

Suppose $\frac{k}{n} \to 0$ and $k \to \infty$, then $\mathbb{E}|f_k(X) - f(X)| \xrightarrow{n \to \infty} 0$

Any function $f$, $\mathbb{E}f^2 < \infty$, no matter how complex.
- $\{X(i)\}_{i=1}^k \rightarrow x$ as long as $k$ is fixed or grows slow $(k/n \rightarrow 0)$
- Suppose $f$ is continuous, then we also get $\{f(X(i))\}_{i=1}^k \rightarrow f(x)$
- If $k \rightarrow \infty$, then $f_k(x) = \frac{1}{k} \sum (f(X(i)) + \text{noise}) \rightarrow f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$'s.
Consider the $k$-NN $\{X_{(i)}\}_{i=1}^k$ of some $x$

- $\{X_{(i)}\}_{i=1}^k \to x$ as long as $k$ is fixed or grows slow ($k/n \to 0$)
- Suppose $f$ is continuous, then we also get $\{f(X_{(i)})\}_{i=1}^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k} \sum (f(X_{(i)}) + \text{noise}) \to f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$'s.
Intuition:

As $n \nearrow$, all $\{X(i)\}_{1}^{k}$ move closer to $x$ as long as $k$ is fixed or grows slow ($k/n \to 0$)

- $\{X(i)\}_{1}^{k} \to x$ as long as $k$ is fixed or grows slow ($k/n \to 0$)
- Suppose $f$ is continuous, then we also get $\{f(X(i))\}_{1}^{k} \to f(x)$
- If $k \to \infty$, then $f_{k}(x) = \frac{1}{k} \sum (f(X(i)) + \text{noise}) \to f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$’s.
Intuition:

As $n \nearrow$, all $\{X(i)\}_1^k$ move closer to $x$ as long as $k$ is fixed or grows slow ($k/n \to 0$)

- $\{X(i)\}_1^k \to x$ as long as $k$ is fixed or grows slow ($k/n \to 0$)
- Suppose $f$ is continuous, then we also get $\{f(X(i))\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k} \sum (f(X(i)) + \text{noise}) \to f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$'s.
Intuition:

As \( n \rightarrow \), all \( \{X(i)\}_1^k \) move closer to \( x \)

- \( \{X(i)\}_1^k \rightarrow x \) as long as \( k \) is fixed or grows slow (\( k/n \rightarrow 0 \))
- Suppose \( f \) is continuous, then we also get \( \{f(X(i))\}_1^k \rightarrow f(x) \)
- If \( k \rightarrow \infty \), then \( f_k(x) = \frac{1}{k} \sum (f(X(i)) + \text{noise}) \rightarrow f(x) \)

Now, any \( f, \mathbb{E}f^2 < \infty \) can be approximated by continuous \( f \)'s.
Intuition:

As $n \to \infty$, all \( \{X(i)\}_1^k \) move closer to $x$

- \( \{X(i)\}_1^k \to x \) as long as $k$ is fixed or grows slow ($k/n \to 0$)
- Suppose $f$ is continuous, then we also get \( \{f(X(i))\}_1^k \to f(x) \)
- If $k \to \infty$, then $f_k(x) = \frac{1}{k} \sum (f(X(i)) + \text{noise}) \to f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$’s.
Intuition:

As $n \nearrow$, all $\{X_{(i)}\}_{1}^{k}$ move closer to $x$

- $\{X_{(i)}\}_{1}^{k} \rightarrow x$ as long as $k$ is fixed or grows slow ($k/n \rightarrow 0$)
- Suppose $f$ is continuous, then we also get $\{f(X_{(i)})\}_{1}^{k} \rightarrow f(x)$
- If $k \rightarrow \infty$, then $f_{k}(x) = \frac{1}{k} \sum(f(X_{(i)}) + \text{noise}) \rightarrow f(x)$

Now, any $f, \mathbb{E}f^2 < \infty$ can be approximated by continuous $f$'s.
Similar universality results for classification, density estimation, ...

**Seminal results on $k$-NN consistency:**

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity, $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.
- [Chaudhuri, Dasgupta, 14]: classification, nice metric/measure.

**Main message:** $k$ should grow (not too fast) with $n$ ... (e.g. $k \sim \log n$)

But we need a more refined picture ...
Similar universality results for classification, density estimation, ...

Seminal results on $k$-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity, $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.
- [Chaudhuri, Dasgupta, 14]: classification, nice metric/measure.

Main message: $k$ should grow (not too fast) with $n$ ... (e.g. $k \sim \log n$)

But we need a more refined picture ...
Similar universality results for classification, density estimation, ...

**Seminal results on \( k \)-NN consistency:**

- [Fix, Hodges, 51]: classification + regularity, \( \mathbb{R}^d \).
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, \( \mathbb{R}^d \).
- [Devroye, Wagner, 77]: density estimation + regularity, \( \mathbb{R}^d \).
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, \( \mathbb{R}^d \).
- [Chaudhuri, Dasgupta, 14]: classification, *nice* metric/measure.

**Main message:** \( k \) should grow (not too fast) with \( n \) ... (e.g. \( k \sim \log n \))

But we need a more refined picture ...
Similar universality results for classification, density estimation, ...

Seminal results on $k$-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity, $\mathbb{R}^d$.
- [Devroye, Gyofri, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.
- [Chaudhuri, Dasgupta, 14]: classification, nice metric/measure.

Main message: $k$ should grow (not too fast) with $n$ ... (e.g. $k \sim \log n$)

But we need a more refined picture ...
Similar universality results for classification, density estimation, ...

Seminal results on $k$-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity, $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.
- [Chaudhuri, Dasgupta, 14]: classification, nice metric/measure.

Main message: $k$ should grow (not too fast) with $n$ ... (e.g. $k \sim \log n$)

But we need a more refined picture ...
PART I: Basic Statistical Insights

- Universality

- **Behavior of** $k$-NN **Distances**

- From Regression to Classification

- Classification is easier than regression

- Multiclass and Mixed Costs
Why $k$-NN Distances?

**Recall Intuition:**
Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ... 

So we hope that $k$-NN($x$) are close to $x$ ... depends on $k$ ... 

Formally: let $r_k(x) \equiv$ distance from $x$ to $k$-th NN in i.i.d. $\{X_i\}_{i=1}^n$ 

Q: How small is $r_k(x) \equiv$ function of $(P_X, k, n)$?
Why $k$-NN Distances?

**Recall Intuition:**

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

So we hope that $k$-NN($x$) are close to $x$ ... depends on $k$ ...

Formally: let $r_k(x) \equiv$ distance from $x$ to $k$-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of $(P_X, k, n)$?
Why $k$-NN Distances?

**Recall Intuition:**

Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

So we hope that $k$-NN($x$) are close to $x$ ... depends on $k$ ...

Formally: let $r_k(x) \equiv$ distance from $x$ to $k$-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of $(P_X, k, n)$?
Why $k$-NN Distances?

**Recall Intuition:**
Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

So we hope that $k$-NN$(x)$ are close to $x$ ... depends on $k$ ...

Formally: let $r_k(x) \equiv$ distance from $x$ to $k$-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of $(P_X, k, n)$?
Why $k$-NN Distances?

**Recall Intuition:**
Closest neighbors of $x$ should be mostly of similar type $y = y(x)$ ...

So we hope that $k$-NN$(x)$ are close to $x$ ... depends on $k$ ...

Formally: let $r_k(x) \equiv$ distance from $x$ to $k$-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of $(P_X, k, n)$?
Fix $x$, and assume \( \{X_i\}_{i=1}^n \) i.i.d. \( P_X \) with density \( p_X \) in \( \mathbb{R}^d \).

\[
B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)
\]

- Assume no ties: \( P_n(B_x) = k/n \).
- w.h.p. \( P_n \approx P_X \implies P_X(B_x) \approx k/n. \)

Now: 
\[
P_X(B_x) \equiv \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d.
\]

Therefore, w.h.p., 
\[
r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}.
\]
Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) = \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{P_X(x)} \cdot \frac{k}{n} \right)^{1/d}$. 
Fix $x$, and assume \( \{X_i\}_{i=1}^n \) i.i.d. \( P_X \) with density \( p_X \) in \( \mathbb{R}^d \).

\[ B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x) \]

- Assume no ties: \( P_n(B_x) = k/n \).
- \text{w.h.p. } P_n \approx P_X \implies P_X(B_x) \approx k/n.

Now: \( P_X(B_x) = \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d \).

Therefore, w.h.p., \( r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \).
Fix $x$, and assume $\{X_i\}_1^n$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) = \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left(\frac{1}{p_X(x) \cdot \frac{k}{n}}\right)^{1/d}$. 
Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing $k$-NN($x$)

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) = \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$. 
Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_{X}$ with density $p_{X}$ in $\mathbb{R}^d$.

$B_{x} \equiv B(x, r_{k}(x)) \equiv$ smallest ball containing $k$-NN($x$)

- Assume no ties: $P_{n}(B_{x}) = \frac{k}{n}$.

- **w.h.p.** $P_{n} \approx P_{X} \implies P_{X}(B_{x}) \approx \frac{k}{n}$.

Now: $P_{X}(B_{x}) = \int_{B_{x}} p_{X}(x') \, dx' \approx p_{X}(x) \cdot \int_{B_{x}} dx' = p_{X}(x) \cdot v_{d} \cdot r_{k}(x)^{d}$.

Therefore, w.h.p., $r_{k}(x) \approx \left( \frac{1}{p_{X}(x)} \cdot \frac{k}{n} \right)^{1/d}$.
Fix $x$, and assume $\{X_i\}_{i=1}^n$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

\[ B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x) \]

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$.
Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing $k$-NN$(x)$

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$. 

Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$. 

$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing $k$-NN$(x)$

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$.
Fix $x$, and assume $\{X_i\}_{1}^{n}$ i.i.d. $P_X$ with density $p_X$ in $\mathbb{R}^d$.

$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- w.h.p. $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x') \, dx' \approx p_X(x) \cdot \int_{B_x} dx' = p_X(x) \cdot v_d \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$.
$$r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$$

Immediate messages:

- $r_k(x) \uparrow$ when local density $p_X(x) \downarrow$
- $r_k(x) \uparrow$ when input dimension $d \uparrow$

Use smaller $k$ for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx \epsilon^{-d}$ ... Fortunately, $d \equiv$ intrinsic dimension($X$) ...
\[ r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \]

**Immediate messages:**

- \( r_k(x) \uparrow \) when local density \( p_X(x) \downarrow \)
- \( r_k(x) \uparrow \) when input dimension \( d \uparrow \)

Use smaller \( k \) for higher dimensional data ...

**Curse of dimension:** For \( r_k \approx \epsilon \) we need \( n \approx \epsilon^{-d} \) ...

Fortunately, \( d \equiv \) intrinsic dimension(\( X \)) ...
\[ r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \]

**Immediate messages:**

- \( r_k(x) \uparrow \) when local density \( p_X(x) \downarrow \)
- \( r_k(x) \uparrow \) when input dimension \( d \uparrow \)

Use smaller \( k \) for higher dimensional data ...

**Curse of dimension:** For \( r_k \approx \epsilon \) we need \( n \approx \epsilon^{-d} \) ...

Fortunately, \( d \equiv \) intrinsic dimension \((X)\) ...
\[ r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \]

**Immediate messages:**

- \( r_k(x) \uparrow \) when local density \( p_X(x) \downarrow \)
- \( r_k(x) \uparrow \) when input dimension \( d \uparrow \)

Use smaller \( k \) for higher dimensional data ...

**Curse of dimension:** For \( r_k \approx \epsilon \) we need \( n \approx \epsilon^{-d} \) ...

Fortunately, \( d \equiv \) intrinsic dimension\((X) \) ...
\[ r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \]

**Immediate messages:**

- \( r_k(x) \uparrow \) when local density \( p_X(x) \downarrow \)
- \( r_k(x) \uparrow \) when input dimension \( d \uparrow \)

**Use smaller \( k \) for higher dimensional data ...**

**Curse of dimension:** For \( r_k \approx \epsilon \) we need \( n \approx \epsilon^{-d} \) ...

Fortunately, \( d \equiv \) intrinsic dimension(\( X \)) ...
\[ r_k(x) \approx \left( \frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d} \]

**Immediate messages:**

- \( r_k(x) \uparrow \) when local density \( p_X(x) \downarrow \)
- \( r_k(x) \uparrow \) when input dimension \( d \uparrow \)
  
  **Use smaller \( k \) for higher dimensional data ...**

**Curse of dimension:** For \( r_k \approx \epsilon \) we need \( n \approx \epsilon^{-d} \) ...

Fortunately, \( d \equiv \) intrinsic dimension(\( X \)) ...

Suppose \( X \in \mathbb{R}^D \), but lies on a \( d \)-dimensional space \( \mathcal{X} \) ...

Consider \( B \), of radius \( r \), centered on \( \mathcal{X} \):

\[
P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d
\]

Thus we’d have \( r_k(x) \approx (k/n)^{1/d} \), irrespective of \( D \gg d \).
Suppose $X \in \mathbb{IR}^D$, but lies on a $d$-dimensional space $\mathcal{X}$. ... 

Consider $B$, of radius $r$, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$. 


Suppose $X \in \mathbb{R}^D$, but lies on a $d$-dimensional space $\mathcal{X}$ ...

Consider $B$, of radius $r$, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we’d have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$. 
Suppose $X \in \mathbb{R}^D$, but lies on a $d$-dimensional space $\mathcal{X}$. …

Consider $B$, of radius $r$, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we’d have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$. 
Suppose $X \in \mathbb{R}^D$, but lies on a $d$-dimensional space $\mathcal{X}$ ...

Consider $B$, of radius $r$, centered on $\mathcal{X}$:

$$P_X(B) \approx p_x \cdot \int_{B \cap \mathcal{X}} dx \approx p_x \cdot r^d$$

Thus we’d have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$. 
Suppose \( X \in \mathbb{R}^D \), but lies on a \( d \)-dimensional space \( \mathcal{X} \) ...

Consider \( B \), of radius \( r \), centered on \( \mathcal{X} \):

\[
P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d
\]

Thus we’d have \( r_k(x) \approx (k/n)^{1/d} \), irrespective of \( D \gg d \).
Quick Simulations:

Embed \((d = 2)\)-data into high-dimensional \(\mathbb{R}^D\), \(D \to \infty\)
Quick Simulations:

Fix \( d = 2 \): average NN distances are stable as \( D \) varies.
Refined analysis for $r_k(x)$:

[J. Costa, A. Hero 04], [R. Samworth 12]

Implications:

$r_k(x)$ adaptive to $d \implies$ NN methods adaptive to $d$ ...

($d$-sparse documents/images, Robotics data on $d$-manifold)
Refined analysis for $r_k(x)$:

[J. Costa, A. Hero 04], [R. Samworth 12]

Implications:

$r_k(x)$ adaptive to $d \implies$ NN methods adaptive to $d$ ...

($d$-sparse documents/images, Robotics data on $d$-manifold)
PART I: Basic Statistical Insights

• Universality

• Behavior of $k$-NN Distances

• **From Regression to Classification**

• Classification is easier than regression

• Multiclass and Mixed Costs
From bounds on \( r_k(x) \) to error rates:

Program:

1. Regression bounds
2. Reduce Classification to Regression
From bounds on $r_k(x)$ to error rates:

Program:

1. Regression bounds
2. Reduce Classification to Regression
**k-NN Regression**

**Data:** \( \{(X_i, Y_i)\}_{i=1}^{n}, \ Y = f(X) + \text{noise} \)

**Learn:** \( f_k(x) = \text{avg } (Y_i) \text{ of } k\text{-NN}(x). \)

**Ideal Metric \( \rho \):** \( f(x) \approx f(x') \text{ if } \rho(x, x') \approx 0 \)

... e.g., assume \( f \) is Lipschitz: \( |f(x) - f(x')| \leq \lambda \cdot \rho(x, x'). \)

**Performance Goal:**

Pick \( k \) such that \( \|f_k - f\|^2 = \mathbb{E}_X |f_k(X) - f(X)|^2 \) is small.
**k-NN Regression**

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, Y = f(X) + \text{noise} \)

Learn: \( f_k(x) = \text{avg } (Y_i) \) of k-NN(x).

**Ideal Metric** \( \rho \): \( f(x) \approx f(x') \) if \( \rho(x, x') \approx 0 \)

... e.g., assume \( f \) is Lipschitz: \( |f(x) - f(x')| \leq \lambda \cdot \rho(x, x') \).

**Performance Goal:**

Pick \( k \) such that \( \|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2 \) is small.
**$k$-NN Regression**

**Data:** $\{(X_i, Y_i)\}_{i=1}^{n}$, $Y = f(X) + \text{noise}$

**Learn:** $f_k(x) = \text{avg } (Y_i) \text{ of } k$-NN$(x)$.

**Ideal Metric $\rho$:** $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume $f$ is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

**Performance Goal:**

Pick $k$ such that $\|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2$ is small.
**$k$-NN Regression**

Data: \( \{(X_i, Y_i)\}_{i=1}^n \), \( Y = f(X) + \text{noise} \)

Learn: \( f_k(x) = \text{avg} \ (Y_i) \) of \( k \)-NN\( (x) \).

**Ideal Metric \( \rho \):** 
\[ f(x) \approx f(x') \text{ if } \rho(x, x') \approx 0 \]

... e.g., assume \( f \) is Lipschitz: 
\[ |f(x) - f(x')| \leq \lambda \cdot \rho(x, x'). \]

**Performance Goal:**

Pick \( k \) such that \( \|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2 \) is small.
**Step 1: Bias-variance decomposition**

**Intuition:** \[ \mathbb{E} |Z - c|^2 = \mathbb{E} |Z - \mathbb{E}Z|^2 + |c - \mathbb{E}Z|^2. \]

So fix \( x \), and fix \( \{X_i\} \), and let \( \tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x) \) ...

\[
\mathbb{E} |f_k(x) - f(x)|^2 = \underbrace{\mathbb{E} |f_k(x) - \tilde{f}_k(x)|^2}_{\text{Variance}} + \underbrace{|f(x) - \tilde{f}_k(x)|^2}_{\text{Bias}^2}.
\]
Step 1: *Bias-variance decomposition*

**Intuition:** \[ \mathbb{E} |Z - c|^2 = \mathbb{E} |Z - \mathbb{E} Z|^2 + |c - \mathbb{E} Z|^2. \]

So fix \( x \), and fix \( \{X_i\} \), and let \( \tilde{f}_k(x) = \mathbb{E}_{Y_i} f_k(x) \) ...

\[ \mathbb{E} |f_k(x) - f(x)|^2 = \underbrace{\mathbb{E} |f_k(x) - \tilde{f}_k(x)|^2}_{\text{Variance}} + \underbrace{|f(x) - \tilde{f}_k(x)|^2}_{\text{Bias}^2}. \]
Step 1: Bias-variance decomposition

Intuition: \[ \mathbb{E} |Z - c|^2 = \mathbb{E} |Z - \mathbb{E} Z|^2 + |c - \mathbb{E} Z|^2. \]

So fix \( x \), and fix \( \{X_i\} \), and let \( \tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x) \) ...

\[ \mathbb{E} |f_k(x) - f(x)|^2 = \mathbb{E} |f_k(x) - \tilde{f}_k(x)|^2 + |f(x) - \tilde{f}_k(x)|^2. \]
**Step 1: Bias-variance decomposition**

**Intuition:** \[ E|Z - c|^2 = E|Z - E_Z|^2 + |c - E_Z|^2. \]

So fix \( x \), and fix \( \{X_i\} \), and let \( \tilde{f}_k(x) = E_{\{Y_i\}} f_k(x) \) ...
**Step 2: Bound the two terms**

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

  $$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

  $$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} |f(X_i) - f(x)|$$

  $$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

  $$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
Step 2: Bound the two terms

- Variance: recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

$$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- Bias: note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

$$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \left| f(X_i) - f(x) \right|$$

$$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

$$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
Step 2: Bound the two terms

- **Variance:** recall \( f_k(x) = \frac{1}{k} \sum_{X_i \in k-NN(x)} Y_i \)

\[
\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-NN(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}
\]

- **Bias:** note that \( \tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-NN(x)} f(X_i) \)

\[
\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-NN(x)} \left| f(X_i) - f(x) \right|
\]
\[
\leq \frac{1}{k} \sum_{X_i \in k-NN(x)} \rho(X_i, x)
\]
\[
\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.
\]
Step 2: Bound the two terms

- **Variance:** recall \( f_k(x) = \frac{1}{k} \sum_{X_i \in \text{k-NN}(x)} Y_i \)

  \[
  \text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in \text{k-NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}
  \]

- **Bias:** note that \( \tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in \text{k-NN}(x)} f(X_i) \)

  \[
  \left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in \text{k-NN}(x)} \left| f(X_i) - f(x) \right| \\
  \leq \frac{1}{k} \sum_{X_i \in \text{k-NN}(x)} \rho(X_i, x) \\
  \leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.
  \]
Step 2: Bound the two terms

- Variance: recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

$$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- Bias: note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

$$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} |f(X_i) - f(x)|$$

$$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

$$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
Step 2: Bound the two terms

- Variance: recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

$$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- Bias: note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

$$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \left| f(X_i) - f(x) \right|$$

$$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

$$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
Step 2: Bound the two terms

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-NN(x)} Y_i$

  $$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-NN(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-NN(x)} f(X_i)$

  $$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-NN(x)} \left| f(X_i) - f(x) \right|$$

  $$\leq \frac{1}{k} \sum_{X_i \in k-NN(x)} \rho(X_i, x)$$

  $$\leq r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.$$
Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

  $$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

  $$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} |f(X_i) - f(x)|$$

  $$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

  $$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
**Step 2:** Bound the two terms

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} Y_i$

  $$\text{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k-\text{NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} f(X_i)$

  $$\left| \tilde{f}_k(x) - f(x) \right| \leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \left| f(X_i) - f(x) \right|$$

  $$\leq \frac{1}{k} \sum_{X_i \in k-\text{NN}(x)} \rho(X_i, x)$$

  $$\leq r_k(x) \approx \left( \frac{k}{n} \right)^{1/d}.$$
Step 3: \textit{Integrate over }x\textit{ and }\{X_i\}

We then get: \[ E\|f_k - f\|^2 \lesssim \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}. \]

Pick \( k = \Theta(n^{2/(2+d)}) \) to get \( E\|f_k - f\|^2 \ll n^{2/(2+d)} \), optimal.

Best choice of \( k \gg n^{2/(2+d)} \) and \( k \ll n \).

Choosing \( k \) by C-V yields same optimal rates. (under reg. on noise)
Step 3: Integrate over $x$ and $\{X_i\}$

We then get: $\mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d}$.

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)}$, optimal.

Best choice of $k \uparrow$ as $n \uparrow$ and $d \downarrow$.

Choosing $k$ by C-V yields same optimal rates. (under reg. on noise)
**Step 3:** \textit{Integrate over }$x$\textit{ and }$\{X_i\}$

We then get: \[ \mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d} . \]

**Tradeoff on }$k$\textit{:}

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)}$, optimal.

**Best choice of }$k$\textit{ as }$n \uparrow$\textit{ and }$d \downarrow$

Choosing $k$ by C-V yields same optimal rates. (under reg. on noise)
Step 3: Integrate over $x$ and $\{X_i\}$

We then get: \[ \mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d}. \]

Tradeoff on $k$:

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)}$, optimal.

Best choice of $k \uparrow$ as $n \uparrow$ and $d \downarrow$.

Choosing $k$ by C-V yields same optimal rates. (under reg. on noise)
Step 3: \textit{Integrate over }x\textit{ and }\{X_i\}

We then get: \[ \mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d}. \]

Tradeoff on \(k\):

Pick \(k = \Theta(n^{2/(2+d)})\) to get \(\mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)}, \) optimal.

Best choice of \(k\) \(\uparrow\) as \(n \uparrow\) and \(d \downarrow\).

Choosing \(k\) by C-V yields same optimal rates. (under reg. on noise)
Step 3: Integrate over $x$ and $\{X_i\}$

We then get:

$$\mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d}.$$  

Tradeoff on $k$:

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)}$, optimal.

Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$.

Choosing $k$ by C-V yields same optimal rates. (under reg. on noise)
Step 3: \textit{Integrate over }x\textit{ and }\{X_i\}

We then get: \[\mathbb{E} \| f_k - f \|^2 \lesssim \frac{1}{k} + \left( \frac{k}{n} \right)^{2/d}.\]

Tradeoff on \( k \):

Pick \( k = \Theta(\frac{n^2}{(2+d)}) \) to get \( \mathbb{E} \| f_k - f \|^2 \lesssim n^{-2/(2+d)} \), optimal.

\textbf{Best choice of } \( k \uparrow \text{ as } n \uparrow \text{ and } d \downarrow \)

Choosing \( k \) by C-V yields same optimal rates. (under reg. on noise)
Similar messages under generalizations of Lipschitz assumption:

- Hölder continuity: \[ |f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha. \]

(avg. version leads to so-called Nikolskii, Sobolev conditions)
(see e.g. [Györfi, Krzyżak, Walk, 02])
Similar messages under generalizations of Lipschitz assumption:

- Hölder continuity: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha$.

(avg. version leads to so-called Nikolskii, Sobolev conditions)

(see e.g. [Györfi, Krzyżak, Walk, 02])
Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** \[ |f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha. \]

  (avg. version leads to so-called Nikolskii, Sobolev conditions)

\[ |x - x'|^\alpha \text{ gets flatter around } x = 0 \text{ as } \alpha \nearrow. \]

(see e.g. [Györfi, Krzyżak, Walk, 02])
Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity**: \[ |f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha. \]
  (avg. version leads to so-called Nikolskii, Sobolev conditions)

\[ |x - x'|^\alpha \text{ gets flatter around } x = 0 \text{ as } \alpha \nearrow. \]

**Additional messages (as } \alpha \nearrow):**

- Local averages (as $k$-NN) not appropriate for smoother (easier) $f$.
- Local polynomials are best, but harder to implement in high-$D$.

(see e.g. [Györfi, Krzyżak, Walk, 02])
Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** \( |f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha. \)
  
  (avg. version leads to so-called Nikolskii, Sobolev conditions)

\[ |x - x'|^\alpha \text{ gets flatter around } x = 0 \text{ as } \alpha \nearrow. \]

**Additional messages (as } \alpha \nearrow):**

- Local averages (as \( k \)-NN) not appropriate for smoother (easier) \( f \).
- Local polynomials are best, but harder to implement in high-\( D \).

(see e.g. [Györfi, Krzyżak, Walk, 02])
From bounds on $r_k(x)$ to error rates:

Program:

1. Regression bounds
2. Reduce Classification to Regression
**k-NN Classification**

**Data:** \( \{(X_i, Y_i)\}_{i=1}^n, \ Y \in \{0, 1\} \).

**Learn:** \( h_k(x) = \text{majority} \ (Y_i) \) of \( k\)-NN\( (x) \).

**Reduces to regression:** let \( f_k(x) = \text{avg} \ (Y_i) \) of \( k\)-NN\( (x) \)

... then: \( h_k(x) \equiv 1 \{f_k(x) \geq 1/2\} \).

**Performance Goal:**

Pick \( k \) such that \( \text{err}(h_k) = \mathbb{P}(h_k(X) \neq Y) \) is small.

Equivalently, consider \( \mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*) \).
**k-NN Classification**

Data: \( \{(X_i, Y_i)\}_{i=1}^n \), \( Y \in \{0, 1\} \).

Learn: \( h_k(x) = \text{majority} (Y_i) \) of \( k \)-NN(x).

Reduces to regression: let \( f_k(x) = \text{avg} (Y_i) \) of \( k \)-NN(x)

... then: \( h_k(x) \equiv 1\{f_k(x) \geq 1/2\} \).

Performance Goal:
Pick \( k \) such that \( \text{err}(h_k) = \mathbb{P}(h_k(X) \neq Y) \) is small.

Equivalently, consider \( \mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*) \).
**k-NN Classification**

Data: \(\{(X_i, Y_i)\}_{i=1}^{n}\), \(Y \in \{0, 1\}\).

Learn: \(h_k(x) = \text{majority}(Y_i) \text{ of } k\text{-NN}(x)\).

Reduces to regression: let \(f_k(x) = \text{avg}(Y_i) \text{ of } k\text{-NN}(x)\)

... then: \(h_k(x) \equiv 1\{f_k(x) \geq 1/2\}\).

Performance Goal:

Pick \(k\) such that \(\text{err}(h_k) \equiv \mathbb{P}(h_k(X) \neq Y)\) is small.

Equivalently, consider \(\mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*)\).
**k-NN Classification**

Data: \( \{(X_i, Y_i)\}_{i=1}^n, Y \in \{0, 1\} \).

Learn: \( h_k(x) = \text{majority} \ (Y_i) \text{ of } k\text{-NN}(x) \).

**Reduces to regression:** let \( f_k(x) = \text{avg} \ (Y_i) \text{ of } k\text{-NN}(x) \)

... then: \( h_k(x) \equiv 1 \{f_k(x) \geq 1/2\} \).

**Performance Goal:**

Pick \( k \) such that \( \text{err}(h_k) \equiv \mathbb{P}(h_k(X) \neq Y) \) is small.

Equivalently, consider \( \mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*) \).
k-NN Classification

Data: \( \{(X_i, Y_i)\}_{i=1}^n \), \( Y \in \{0, 1\} \).

Learn: \( h_k(x) = \text{majority} (Y_i) \) of \( k\)-NN\( (x) \).

Reduces to regression: let \( f_k(x) = \text{avg} (Y_i) \) of \( k\)-NN\( (x) \)

... then: \( h_k(x) \equiv 1 \{ f_k(x) \geq 1/2 \} \).

Performance Goal:

Pick \( k \) such that \( \text{err}(h_k) \equiv P(h_k(X) \neq Y) \) is small.

Equivalently, consider \( \mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*) \).
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = P(Y = 1|x)$, and $h^*(x) = 1\{f(x) \geq 1/2\}$, while $h_k(x) = 1\{f_k(x) \geq 1/2\}$.

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| dP_X \leq 2\|f_k - f\|$. 

For Lipschitz $f$: $\mathbb{E}\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$. 

Similar messages on choice of $k$: ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$, and $h^*(x) = 1\{f(x) \geq 1/2\}$, while $h_k(x) = 1\{f_k(x) \geq 1/2\}$.

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2|dP_X \leq 2\|f_k - f\|$.

For Lipschitz $f$: $\mathbb{E}\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of $k$: ...
Remarks: \( f_k(x) \) estimates \( f(x) \equiv \mathbb{E} [Y | x] = \mathbb{P}(Y = 1 | x) \), and \( h^*(x) = \mathbb{1} \{ f(x) \geq 1/2 \} \), while \( h_k(x) \equiv \mathbb{1} \{ f_k(x) \geq 1/2 \} \).

\[ f_k(x) \approx f(x) \text{ implies } h_k(x) = h^*(x) \]

Formally: \[ \mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| \, dP_x \leq 2 \| f_k - f \|. \]

For Lipschitz \( f \): \( \mathbb{E} \mathcal{E}(h_k) \lesssim n^{-1/(2+d)} \), for \( k = \Theta(n^{2/(2+d)}) \).

Similar messages on choice of \( k \): ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$, and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| \, dP_X \leq 2 \|f_k - f\|$. 

For Lipschitz $f$: $\mathbb{E}\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$. Similar messages on choice of $k$: ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E} [Y| x] = \mathbb{P}(Y = 1|x)$, and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| \, dP_X \leq 2 \|f_k - f\|.$

For Lipschitz $f$: $\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of $k$: ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$, and $h^*(x) = 1\{f(x) \geq 1/2\}$, while $h_k(x) \equiv 1\{f_k(x) \geq 1/2\}$.

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| \, dP_x \leq 2 \|f_k - f\|$. For Lipschitz $f$: $\mathbb{E}\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$. Similar messages on choice of $k$: ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E} [Y | x] = \mathbb{P}(Y = 1 | x)$, and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Formally:

$$
\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| \, dP_X \leq 2 \|f_k - f\|.
$$

For Lipschitz $f$:

$$
\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-1/(2+d)}, \text{ for } k = \Theta(n^{2/(2+d)}).
$$

Similar messages on choice of $k$: ...
Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$, and $h^*(x) = 1\{f(x) \geq 1/2\}$, while $h_k(x) \equiv 1\{f_k(x) \geq 1/2\}$.

Formally: $\mathcal{E}(h_k) = \int_{h_k \neq h^*} 2|f(x) - 1/2| dP_X \leq 2\|f_k - f\|$. 

For Lipschitz $f$: $\mathbb{E}\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$. 

Similar messages on choice of $k$ ...
PART I: Basic Statistical Insights

- Universality
- Behavior of $k$-NN Distances
- From Regression to Classification
- **Classification is easier than regression**
- Multiclass and Mixed Costs
\( h^* = \mathbb{1}\{f \geq 1/2\}, \text{ while } h_k = \mathbb{1}\{f_k \geq 1/2\}. \)

Suppose \(|f(x) - 1/2| \geq \delta\) for most values \(x\) ... Then \(|f_k - f| \leq \delta\) implies \(h_k = h^*\) often ... no need for \(f_k \approx f\).

**Tsybakov’s noise condition:** \( \mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta \)

If \(|f_k - f| \leq \delta_n\), then \(\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta. \)

For Lipschitz \(f\): \(\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}, \text{ for } k = \Theta(n^{2/(2+d)}). \)
\[ h^* = \mathbb{1}\{f \geq 1/2\}, \text{ while } h_k \equiv \mathbb{1}\{f_k \geq 1/2\}. \]

Suppose \(|f(x) - 1/2| \geq \delta\) for most values \(x\) …

Then \(|f_k - f| \leq \delta\) implies \(h_k = h^*\) often … no need for \(f_k \approx f\).

**Tsybakov’s noise condition:** \(\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta\)

If \(|f_k - f| \leq \delta_n\), then \(\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta\).

For Lipschitz \(f\): \[ \mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}, \text{ for } k = \Theta(n^2/(2+d)). \]
\[ h^* = \mathbb{1}\{f \geq 1/2\}, \text{ while } h_k \equiv \mathbb{1}\{f_k \geq 1/2\}. \]

Suppose \(|f(x) - 1/2| \geq \delta\) for most values \(x\) ...

Then \(|f_k - f| \leq \delta\) implies \(h_k = h^*\) often ... no need for \(f_k \approx f\).

**Tsybakov’s noise condition:** \(\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta\)

If \(|f_k - f| \leq \delta_n\), then \(\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta\).

For Lipschitz \(f\): \(\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}\), for \(k = \Theta(n^{2/(2+d)})\).
Suppose \(|f(x) - 1/2| \geq \delta\) for most values \(x\) ...

Then \(|f_k - f| \leq \delta\) implies \(h_k = h^*\) often ... no need for \(f_k \approx f\).

**Tsybakov’s noise condition:** \(\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta\)

If \(|f_k - f| \leq \delta_n\), then \(\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta\).

For Lipschitz \(f\): \(\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}\), for \(k = \Theta(n^{2/(2+d)})\).
For Lipschitz $f$: $\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of $f$, and intrinsic $d$.
- Large margin $\beta$ mitigates effects of metric:
  e.g. $\beta \geq d/2 \implies$ rate of $n^{-1/2}$ (no curse of dimension!)

Technical Remarks:
- Above rates assume $P_X \equiv$ Uniform.
  ([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform $P_X$, rates get worse, but understudied.
  ([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).
For Lipschitz $f$: \[ \mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}, \] for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of $f$, and intrinsic $d$.
- Large margin $\beta$ mitigates effects of metric:
  e.g. $\beta \geq d/2 \implies$ rate of $n^{-1/2}$ (no curse of dimension!)

Technical Remarks:
- Above rates assume $P_X \equiv \text{Uniform}$. ([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform $P_X$, rates get worse, but understudied. ([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).
For Lipschitz $f$: \( \mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)} \), for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of $f$, and intrinsic $d$.
- Large margin $\beta$ mitigates effects of metric:
  - e.g. $\beta \geq d/2 \implies$ rate of $n^{-1/2}$ (no curse of dimension!)

Technical Remarks:
- Above rates assume $P_X \equiv$ Uniform.
  ([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform $P_X$, rates get worse, but understudied.
  ([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).
For Lipschitz $f$:\n\[ \mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}, \text{ for } k = \Theta(n^{2/(2+d)}). \]

- Choice of metric $\rightarrow$ Lipschitzness of $f$, and intrinsic $d$.
- Large margin $\beta$ mitigates effects of metric:
  e.g. $\beta \geq d/2 \implies$ rate of $n^{-1/2}$ (no curse of dimension!)

Technical Remarks:
- Above rates assume $P_X \equiv \text{Uniform}$.
  ([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform $P_X$, rates get worse, but understudied.
  ([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).
For Lipschitz $f$: $\mathbb{E} \mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of $f$, and intrinsic $d$.
- Large margin $\beta$ mitigates effects of metric:
  e.g. $\beta \geq d/2 \implies$ rate of $n^{-1/2}$ (no curse of dimension!)

**Technical Remarks:**
- Above rates assume $P_X \equiv$ Uniform.
  ([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform $P_X$, rates get worse, but understudied.
  ([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).
PART I: Basic Statistical Insights

- Universality
- Behavior of $k$-NN Distances
- From Regression to Classification
- Classification is easier than regression
- Multiclass and Mixed Costs
Data: \( \{(X_i, Y_i)\}_{i=1}^{n} \), \( Y \in \{1, \ldots, L\} \).

Learn: \( h_k(x) = \) majority \( (Y_i) \) of \( k\text{-NN}(x) \).

Reduction: let \( f_y^k(x) = \) proportion \( (Y = y) \) out of \( k\text{-NN}(x) \).

It estimates \( f^y(x) = \Pr(Y = y|x) \).

... then: \( h_k(x) \equiv \arg\max_y \{f_y^k(x)\} \), and \( h^*(x) = \arg\max_y \{f^y(x)\} \).

Previous insights naturally extends to multiclass ...
$k$-NN extends naturally to multiclass

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, \ldots, L\}$.

Learn: $h_k(x) = \text{majority } (Y_i) \text{ of } k$-NN$(x)$.

Reduction: let $f^y_k(x) = \text{proportion } (Y = y) \text{ out of } k$-NN$(x)$.
It estimates $f^y(x) = P(Y = y|x)$.

... then: $h_k(x) \equiv \arg\max_y \{f^y_k(x)\}$, and $h^*(x) = \arg\max_y \{f^y(x)\}$

Previous insights naturally extends to multiclass ...
**k-NN extends naturally to multiclass**

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, \ Y \in \{1, \ldots, L\} \).

Learn: \( h_k(x) = \text{majority} (Y_i) \text{ of } k-\text{NN}(x) \).

**Reduction:** let \( f_y^k(x) = \text{proportion} (Y = y) \text{ out of } k-\text{NN}(x) \).

It estimates \( f_y(x) = \mathbb{P}(Y = y | x) \).

... then: \( h_k(x) \equiv \arg\max_y \{f_y^k(x)\} \), and \( h^*(x) = \arg\max_y \{f_y(x)\} \).

Previous insights naturally extends to multiclass ...
*k-NN extends naturally to multiclass*

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, Y \in \{1, \ldots, L\} \).

Learn: \( h_k(x) = \text{majority} (Y_i) \text{ of } k-\text{NN}(x) \).

**Reduction:** let \( f_k^y(x) = \text{proportion} (Y = y) \text{ out of } k-\text{NN}(x) \)

It estimates \( f_k^y(x) = P(Y = y|x) \).

... then: \( h_k(x) \equiv \text{argmax}_y \{ f_k^y(x) \}, \text{ and } h^*(x) = \text{argmax}_y \{ f^y(x) \} \)

Previous insights naturally extends to multiclass ...
*k*-NN extends naturally to multiclass

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, Y \in \{1, \ldots, L\} \).

Learn: \( h_k(x) = \text{majority} (Y_i) \) of \( k\text{-NN}(x) \).

Reduction: let \( f^y_k(x) = \text{proportion} (Y = y) \) out of \( k\text{-NN}(x) \).
It estimates \( f^y(x) = \mathbb{P}(Y = y|x) \).

... then: \( h_k(x) \equiv \text{argmax}_y \{f^y_k(x)\}, \text{ and } h^*(x) = \text{argmax}_y \{f^y(x)\} \)

Previous insights naturally extends to multiclass ...
• **Lipschitzness:** \( \| f(x) - f(x') \| \leq \rho(x, x') \)

• **Noise margin:** At any \( x \), we want \( f^{(1)}(x) \gg f^{(2)}(x) \) ...

  assume \( \mathbb{P}_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta \)

Then: \( \mathbb{E} \mathcal{E}(h_k) \lesssim (n/\log L)^{-(\beta+1)/(2+d)}, \) for \( k = \Theta(n^{2/(2+d)}) \).

  Same messages as earlier ...
• **Lipschitzness:** \[ \| f(x) - f(x') \| \leq \rho(x, x') \]

• **Noise margin:** At any \( x \), we want \( f^{(1)}(x) \gg f^{(2)}(x) \) ...

  assume \[ \mathbb{P}_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta \]

Then: \[ \mathbb{E} \mathcal{E}(h_k) \lesssim \left( \frac{n}{\log L} \right)^{-\frac{(\beta+1)}{(2+d)}}, \text{ for } k = \Theta\left( \frac{n^2}{(2+d)} \right). \]

    Same messages as earlier ...
• **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

• **Noise margin:** At any $x$, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ... assume $P_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta$

Then: $\mathbb{E} \mathcal{E}(h_k) \lesssim (n/ \log L)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.  

Same messages as earlier ...
• **Lipschitzness:** \( \| f(x) - f(x') \| \leq \rho(x, x') \)

• **Noise margin:** At any \( x \), we want \( f^{(1)}(x) \gg f^{(2)}(x) \) ...
  
  **assume** \( \mathbb{P}_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta \)

**Then:** \( \mathbb{E} \mathcal{E}(h_k) \lesssim \frac{n}{\log L} \frac{1}{(\beta+1)/(2+d)} \), for \( k = \Theta(n^{2/(2+d)}) \).

Same messages as earlier ...
• **Lipschitzness:** \[ \| f(x) - f(x') \| \leq \rho(x, x') \]

• **Noise margin:** At any \( x \), we want \( f^{(1)}(x) \gg f^{(2)}(x) \) ...

  assume \[ \mathbb{P}_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta \]

Then: \[ \mathbb{E} \mathcal{E}(h_k) \lesssim \left( \frac{n}{\log L} \right)^{-(\beta+1)/(2+d)}, \text{ for } k = \Theta\left( n^{2/(2+d)} \right) \].

Same messages as earlier ...
• **Lipschitzness:** \[ \| f(x) - f(x') \| \leq \rho(x, x') \]

• **Noise margin:** At any \( x \), we want \( f^{(1)}(x) \gg f^{(2)}(x) \) ...
  
  assume \( \mathbb{P}_X \left( f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta \)

Then: \[ \mathbb{E} \mathcal{E}(h_k) \lesssim \left( \frac{n}{\log L} \right)^{-(\beta+1)/(2+d)}, \text{ for } k = \Theta\left( \frac{n^2}{(2+d)} \right). \]

Same messages as earlier ...
Mostly open:

**Mixed costs regimes (e.g., medicine, finance, ...)**

\[ y \leftrightarrow \text{Expected cost when } y \text{ is wrong} \neq 1 - P(Y = y) \]

Natural extensions of previous insights considered in [Reeve, Brown 17]

**Practical Q:** Estimating costs in practice, and integrating with NN ...

Performance metric Elicitation in [S. Koyejo et.al. 19]
Mostly open:

**Mixed costs regimes (e.g., medicine, finance, ...)**

\[ y \rightarrow \text{Expected cost when } y \text{ is wrong } \neq 1 - P(Y = y) \]

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Q: Estimating costs in practice, and integrating with NN ...

Performance metric Elicitation in [S. Koyejo et.al. 19]
Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$y \leftarrow$ Expected cost when $y$ is wrong $\neq 1 - P(Y = y)$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Q: Estimating costs in practice, and integrating with NN ...

Performance metric Elicitation in [S. Koyejo et.al. 19]
Mostly open:

**Mixed costs regimes (e.g., medicine, finance, ...)**

\[ y \leftarrow \text{Expected cost when } y \text{ is wrong} \neq 1 - P(Y = y) \]

Natural extensions of previous insights considered in [Reeve, Brown 17]

**Practical Q:** Estimating costs in practice, and integrating with NN ...

Performance metric Elicitation in [S. Koyejo et.al. 19]
Mostly open:

**Mixed costs regimes (e.g., medicine, finance, ...)**

\[ y \leftarrow \text{Expected cost when } y \text{ is wrong} \neq 1 - P(Y = y) \]

Natural extensions of previous insights considered in [Reeve, Brown 17]

**Practical Q:** Estimating costs in practice, and integrating with NN ...

Performance metric Elicitation in [S. Koyejo et.al. 19]
Important Remark:

Continuity of $f(x) = \mathbb{E}[Y|x]$ is unnatural in classification ...

- **Piecewise Smoothness:** Ben-David, Scott, Nowak, Castro ...
  ($k$-NN not well-understood in these settings)

- **Volume-based smoothness:** [Chaudhuri, Dasgupta 14]

  $$|f(B(x,r)) - f(x)| \leq P_X(B(x,r))^{\alpha/d}$$
Important Remark:

Continuity of $f(x) = \mathbb{E}[Y|x]$ is unnatural in classification ...

• **Piecewise Smoothness**: Ben-David, Scott, Nowak, Castro ...
  ($k$-NN not well-understood in these settings)

• **Volume-based smoothness**: [Chaudhuri, Dasgupta 14]

\[ |f(B(x,r)) - f(x)| \leq P_X(B(x,r))^{\alpha/d} \]
Important Remark:

Continuity of \( f(x) = \mathbb{E}[Y|x] \) is unnatural in classification ...

- **Piecewise Smoothness:** Ben-David, Scott, Nowak, Castro ...
  \((k\text{-NN not well-understood in these settings})\)

- **Volume-based smoothness:** [Chaudhuri, Dasgupta 14]

\[
|f(B(x, r)) - f(x)| \leq P_X(B(x, r))^{\alpha/d}
\]
End of Part I