Online Consistency of the Nearest Neighbor Rule

Chicago Junior Theorists Workshop

Sanjoy Dasgupta and Geelon So
December 6, 2024

Weather prediction problem

On each day $n = 1, 2, \ldots$

- Obtain weather measurements/signals X_n

Weather prediction problem

On each day $n = 1, 2, \ldots$

- Obtain weather measurements/signals X_n
- Predict whether it will rain or shine the next day $\hat{Y}_n$

Weather prediction problem

On each day $n = 1, 2, \ldots$

- Obtain weather measurements/signals X_n
- Predict whether it will rain or shine the next day $\hat{Y}_n$
- Observe ground-truth outcome Y_n

What's a good prediction rule?

A prediction rule decides how *past experiences* are incorporated into *future predictions.*

- We would like predictions to improve over time.

One qualitative notion of learning

Definition (Consistency)

A prediction rule is consistent if its mistake rate vanishes:

$$\limsup_{N\to\infty} \underbrace{\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{\hat{Y}_n \neq Y_n\}}_{\textit{average mistake rate}} = 0.$$

One qualitative notion of learning

Definition (Consistency)

A prediction rule is consistent if its mistake rate vanishes:

$$\limsup_{N\to\infty} \underbrace{\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{\hat{Y}_n \neq Y_n\}}_{\textit{average mistake rate}} = 0.$$

- Let's work in the realizable setting, in which making no mistakes is possible.
- Perhaps X_n is a "sufficient set" of signals so that $Y_n = \eta(X_n)$ is a function of X_n.

When is consistency possible?

- Conversely, what makes learning hard?

Classical results

Realizable online classification ("Littlestone setting")

- The sequence $\mathbb{X} = (X_n)_n$ may be arbitrary/worst-case.
- Learning requires strong inductive biases on η.

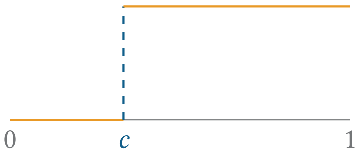
Theorem (Littlestone (1988); Bousquet et al. (2021); etc.)

Consistency is possible $\iff$ *there are only finitely many things to learn about* η*.*

Classical results

Example (Threshold functions are not online learnable)

Let $\mathcal{F}_{\text{threshold}}$ *be the class of threshold functions on the unit interval* $\mathcal{X} = [0, 1]$.

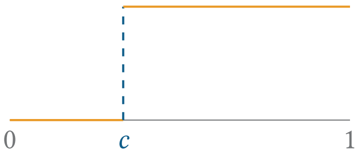


$$\eta_c(x) = \mathbb{1}\{x \geq c\}$$

Classical results

Example (Threshold functions are not online learnable)

Let $\mathcal{F}_{\text{threshold}}$ be the class of threshold functions on the unit interval $\mathcal{X} = [0, 1]$.



$$\eta_c(x) = \mathbb{1}\{x \geq c\}$$

- There is no consistent learner for this function class over arbitrary sequences.
- Informal reason: specifying c requires infinite precision.

Learning in the worst-case setting is hard

- Even with very strong and correct inductive biases, consistency may be impossible.
- The vast majority of learning theory works in settings where learning is ‘easy’.

Classical results

Statistical learning (i.i.d. setting)

- Strong statistical assumption imposed on $\mathbb{X} = (X_n)_n$ such as $X_n \overset{\text{i.i.d.}}{\sim} \nu$.
- Learning is possible even over the class of all measurable functions η.

Theorem (Devroye et al. (2013); Bousquet et al. (2021); etc.)

There are consistent learners in the statistical learning setting.

What made the statistical setting easier?

- And, what sort of trade offs can be made between the hardness of $\mathbb{X}$ and η?

Trading off between sequence class and function class

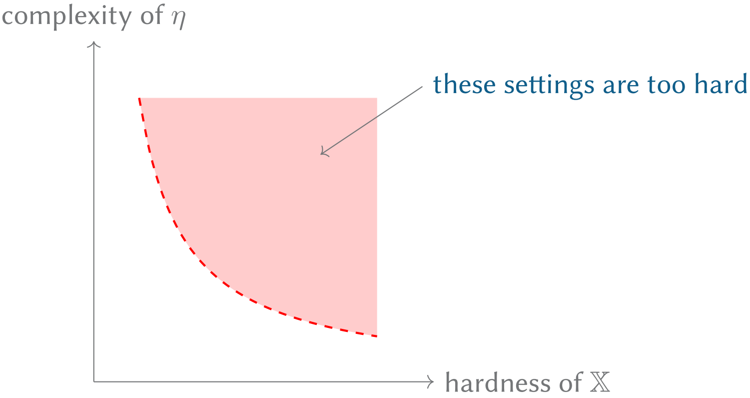


Figure 1: Classical results have largely focused on the extremal settings. Far less is known about what happens in between.

Where does the weather prediction problem fall?

- Weather does not seem to be an i.i.d. nor worst-case phenomenon.
- Learning to predict the weather does not seem to be impossible.

Learning under non-worst case conditions

Figure 2: As classical learning theory often does not capture learning settings in practice, this has motivated the area of non-worst case analysis or smoothed analysis of online learning.

This talk

- Discuss learning through the lens of the nearest neighbor rule
- Introduce some classes of non-worst case sequences and their trade offs

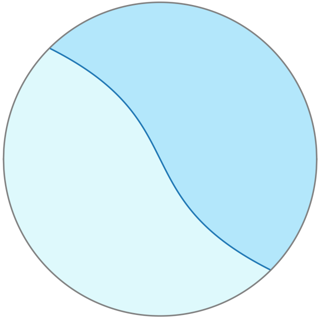
Outline of remainder of talk

1. The nearest neighbor rule
2. Consistency on nice functions
3. Consistency on all functions
4. Takeaways and open problems

The nearest neighbor rule

The realizable online setting

Setup. Let $\mathcal{X}$ be an instance space and $\mathcal{Y}$ be a finite label space. Let $\eta : \mathcal{X} \to \mathcal{Y}$ be the target classifier.



Online classification loop.
For $n = 1, 2, \ldots$

- A test instance X_n is generated.
- The learner makes prediction $\hat{Y}_n$.
- The answer $Y_n = \eta(X_n)$ is revealed.

Consistency of learner: $$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{\hat{Y}_n \neq Y_n\} = 0.$$

The nearest neighbor rule Fix and Hodges (1951)

- Memorize all data points as they come.

The nearest neighbor rule Fix and Hodges (1951)

- Memorize all data points as they come.
- Predict using the label of the most similar instance in memory.

Nearest neighbor process

Let $\mathbb{X} = (X_n)_{n \geq 0}$ be a process on a metric space $(\mathcal{X}, \rho)$.

Nearest neighbor process

Let $\mathbb{X} = (X_n)_{n\geq 0}$ be a process on a metric space $(\mathcal{X}, \rho)$.

Definition

*A **nearest neighbor process** is a sequence $\tilde{\mathbb{X}} = (\tilde{X}_n)_{n>0}$ satisfying*

$$\tilde{X}_n = \operatorname*{arg\,min}_{x \in \mathbb{X}_{<n}} \rho(X_n, x).$$

Nearest neighbor process

Let $\mathbb{X} = (X_n)_{n \geq 0}$ be a process on a metric space $(\mathcal{X}, \rho)$.

Definition

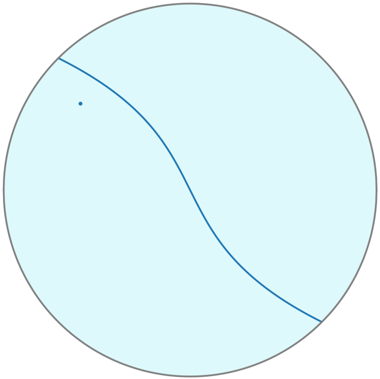
A nearest neighbor process is a sequence $\tilde{\mathbb{X}} = (\tilde{X}_n)_{n>0}$ satisfying

$$\tilde{X}_n = \arg\min_{x \in \mathbb{X}_{<n}} \rho(X_n, x).$$

- The nearest neighbor rule: $\hat{Y}_n = \eta(\tilde{X}_n)$.

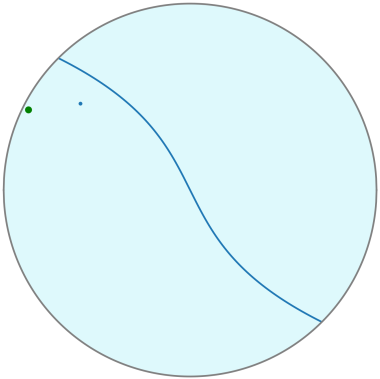
Behavior of the nearest neighbor rule in the i.i.d. setting.

I.I.D. sequence



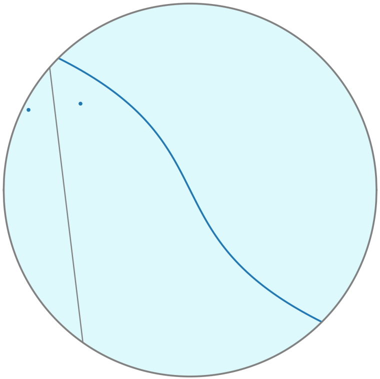
Time	0
Mistake counter	0

I.I.D. sequence



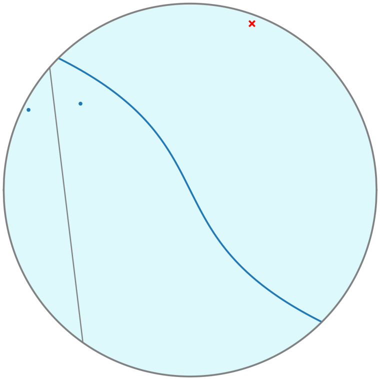
Time	1
Mistake counter	0

I.I.D. sequence



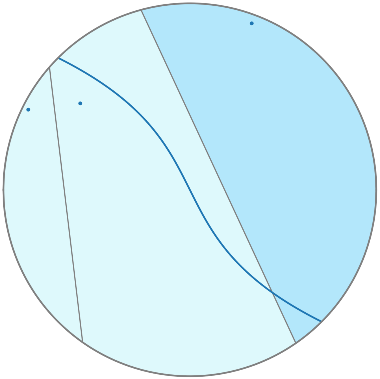
Time	1
Mistake counter	0

I.I.D. sequence



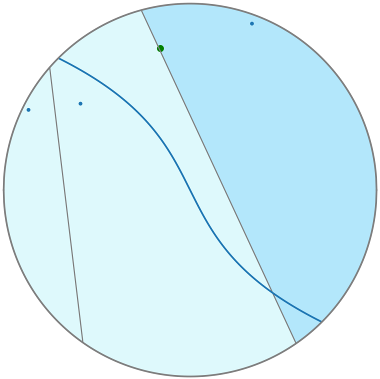
Time	2
Mistake counter	1

I.I.D. sequence



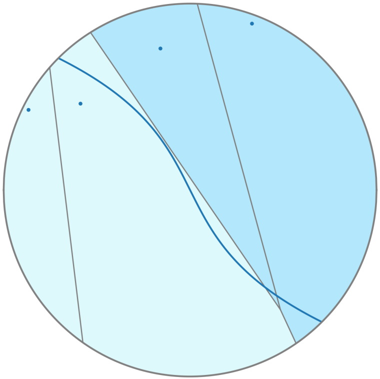
Time	2
Mistake counter	1

I.I.D. sequence



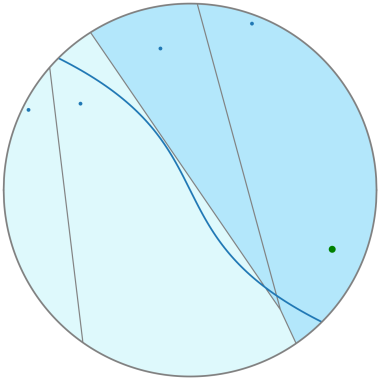
Time	3
Mistake counter	1

I.I.D. sequence



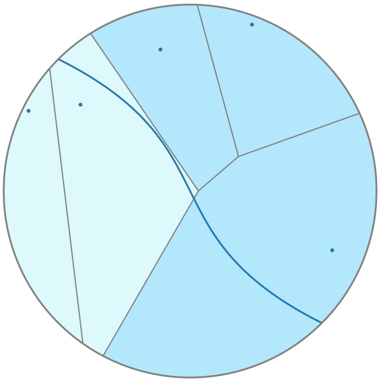
Time	3
Mistake counter	1

I.I.D. sequence



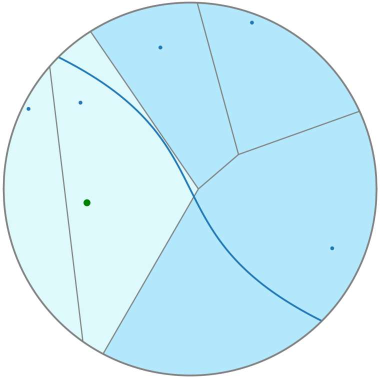
Time	4
Mistake counter	1

I.I.D. sequence



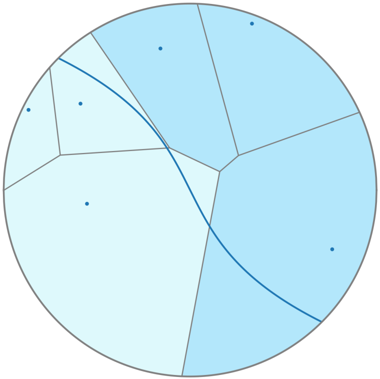
Time	4
Mistake counter	1

I.I.D. sequence



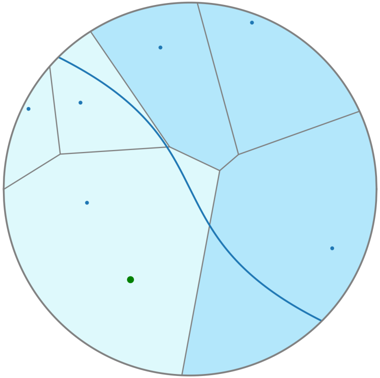
Time	5
Mistake counter	1

I.I.D. sequence



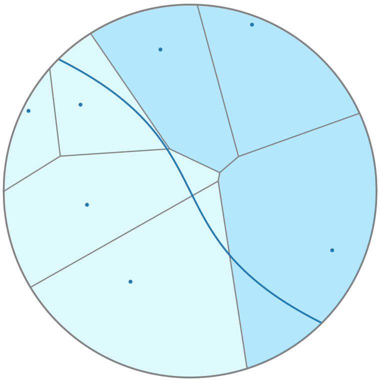
Time	5
Mistake counter	1

I.I.D. sequence



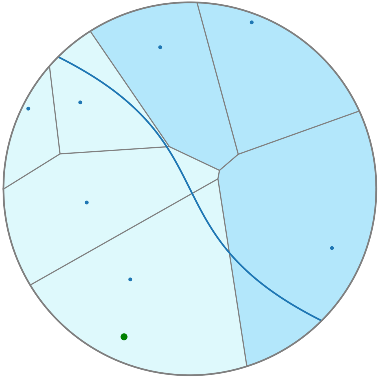
Time	6
Mistake counter	1

I.I.D. sequence



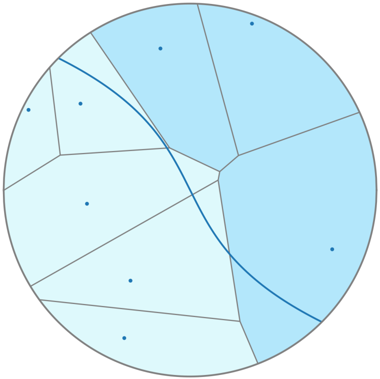
Time	6
Mistake counter	1

I.I.D. sequence



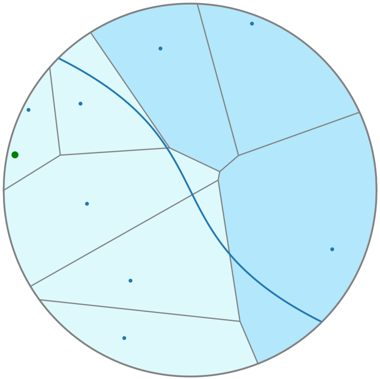
Time	7
Mistake counter	1

I.I.D. sequence



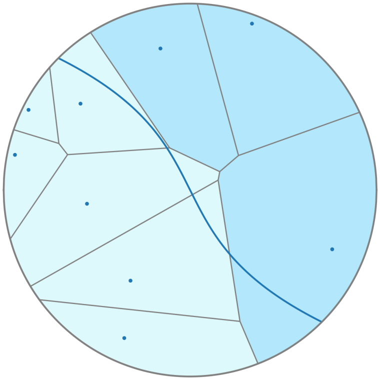
Time	7
Mistake counter	1

I.I.D. sequence



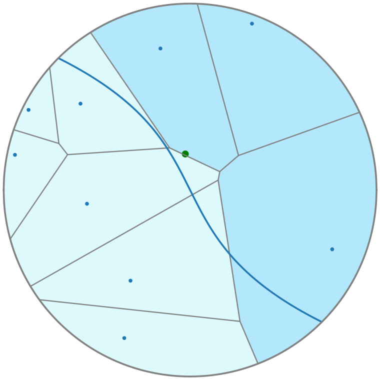
Time	8
Mistake counter	1

I.I.D. sequence



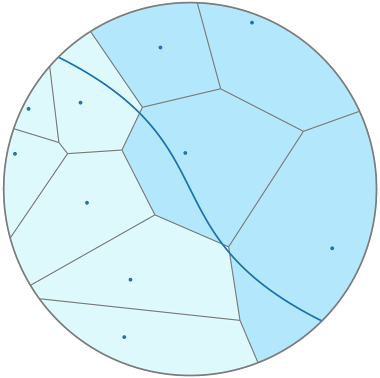
Time	8
Mistake counter	1

I.I.D. sequence



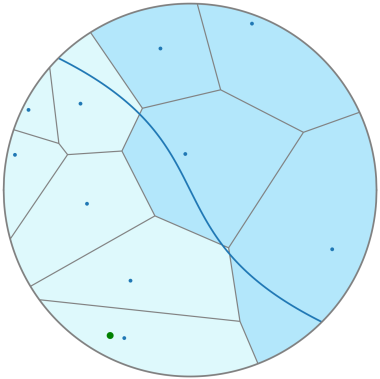
Time	9
Mistake counter	1

I.I.D. sequence



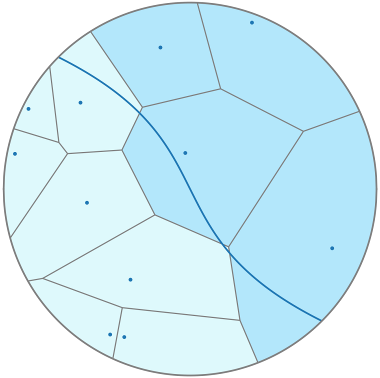
Time	9
Mistake counter	1

I.I.D. sequence



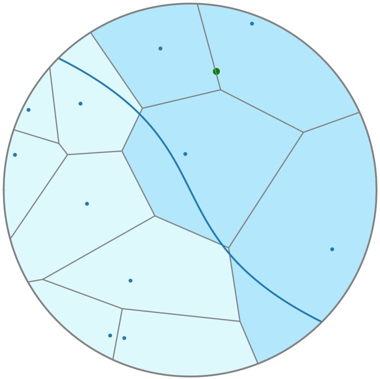
Time	10
Mistake counter	1

I.I.D. sequence



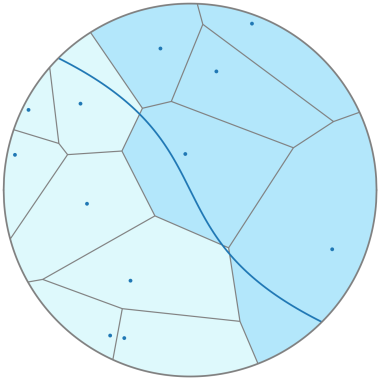
Time	10
Mistake counter	1

I.I.D. sequence



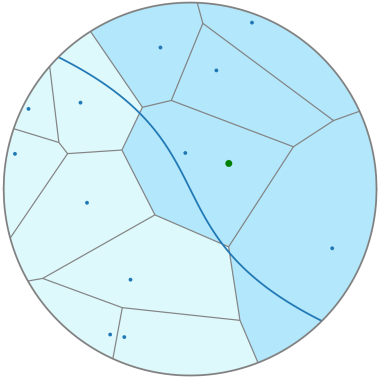
Time	11
Mistake counter	1

I.I.D. sequence



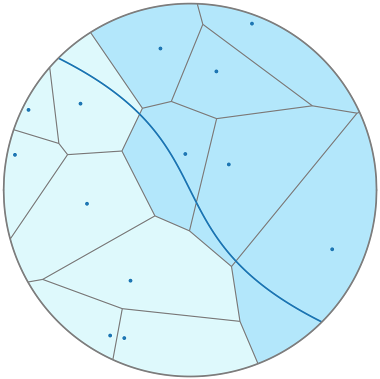
Time	11
Mistake counter	1

I.I.D. sequence



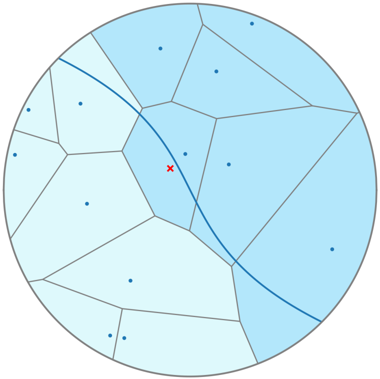
Time	12
Mistake counter	1

I.I.D. sequence



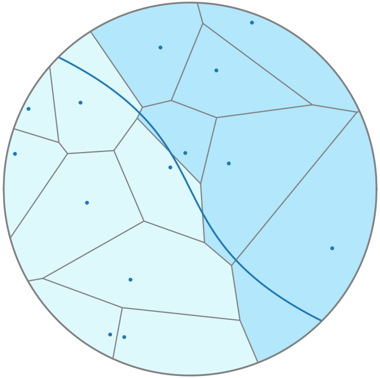
Time	12
Mistake counter	1

I.I.D. sequence



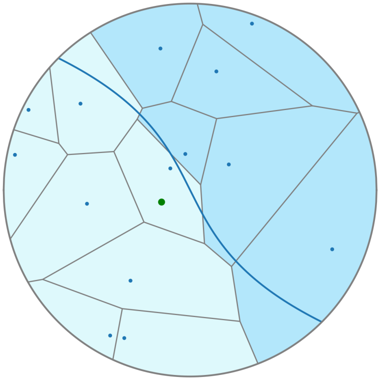
Time	13
Mistake counter	2

I.I.D. sequence



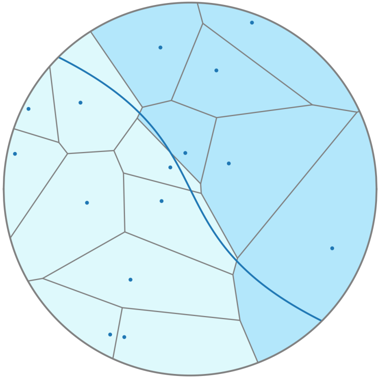
Time	13
Mistake counter	2

I.I.D. sequence



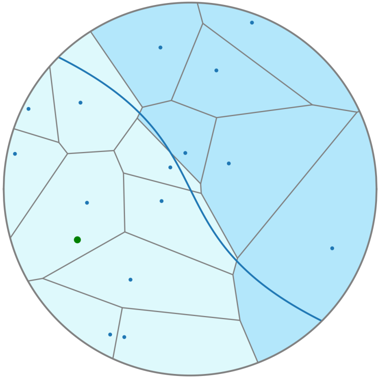
Time	14
Mistake counter	2

I.I.D. sequence



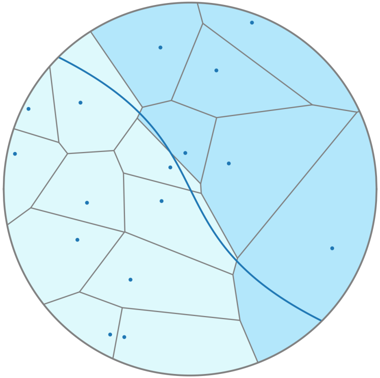
Time	14
Mistake counter	2

I.I.D. sequence



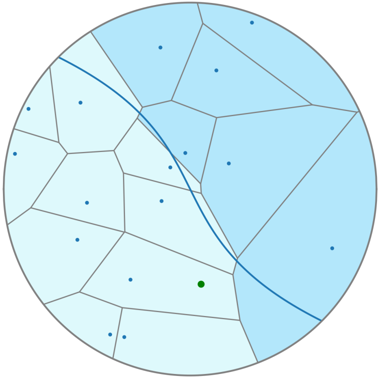
Time	15
Mistake counter	2

I.I.D. sequence



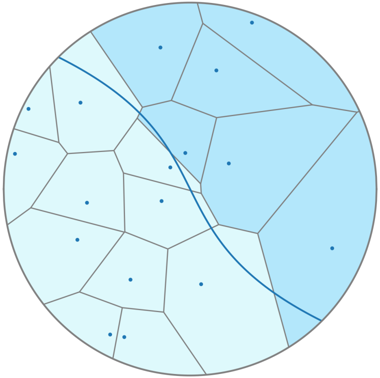
Time	15
Mistake counter	2

I.I.D. sequence



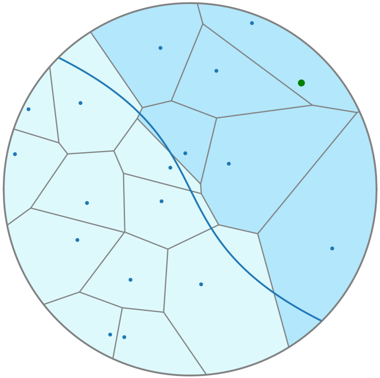
Time	16
Mistake counter	2

I.I.D. sequence



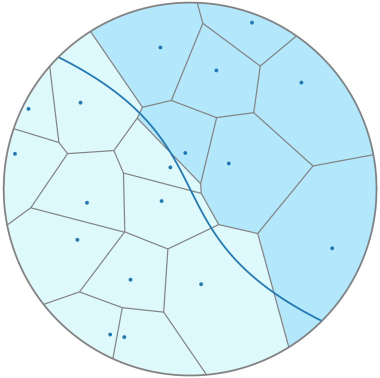
Time	16
Mistake counter	2

I.I.D. sequence



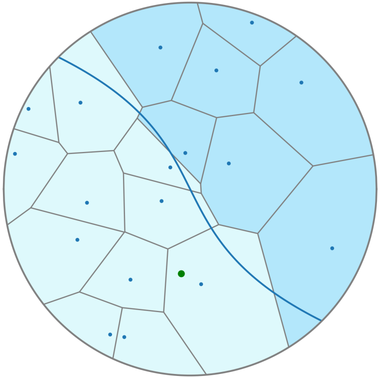
Time	17
Mistake counter	2

I.I.D. sequence



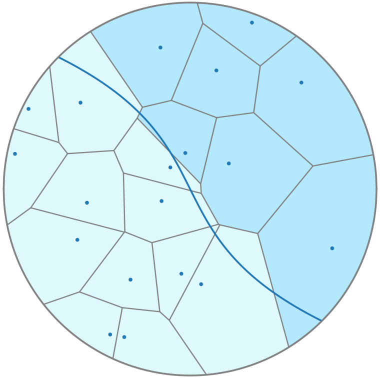
Time	17
Mistake counter	2

I.I.D. sequence



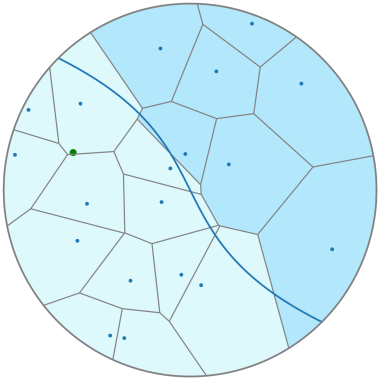
Time	18
Mistake counter	2

I.I.D. sequence



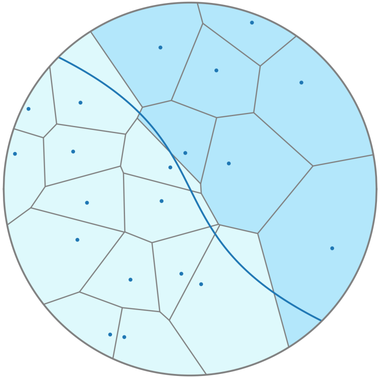
Time	18
Mistake counter	2

I.I.D. sequence



Time	19
Mistake counter	2

I.I.D. sequence



Time	19
Mistake counter	2

Consistent settings for 1-nearest neighbor

i.i.d. — arbitrary

Cover and Hart (1967)

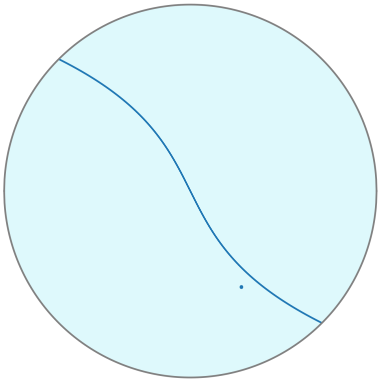
η has negligible boundary

$\mathcal{X}$ is separable

Behavior of the nearest neighbor rule in the worst-case setting.

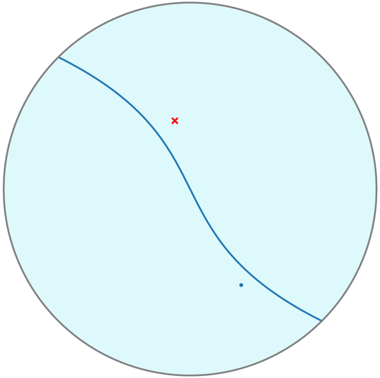
Worst-case sequence

Time	0
Mistake counter	0



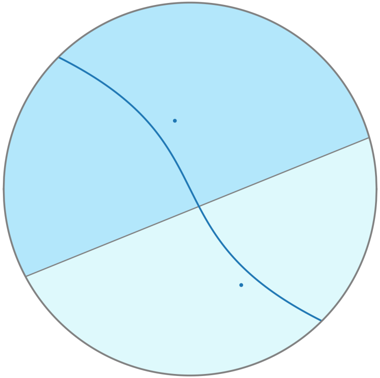
Worst-case sequence

Time	1
Mistake counter	1



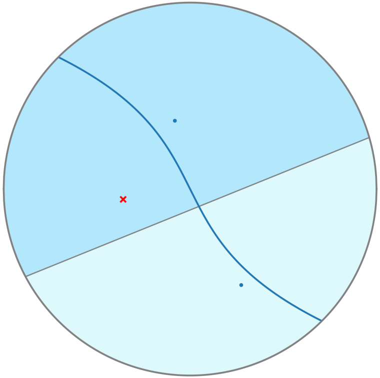
Worst-case sequence

Time	1
Mistake counter	1



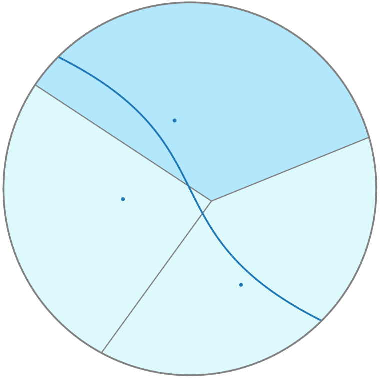
Worst-case sequence

Time	2
Mistake counter	2



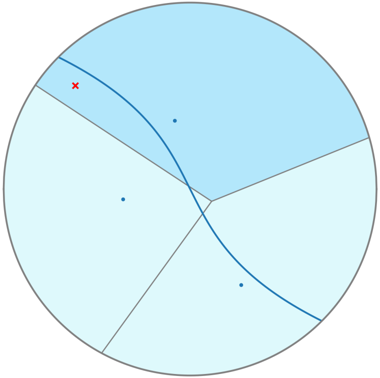
Worst-case sequence

Time	2
Mistake counter	2



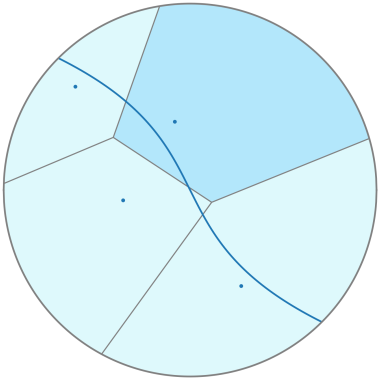
Worst-case sequence

Time	3
Mistake counter	3



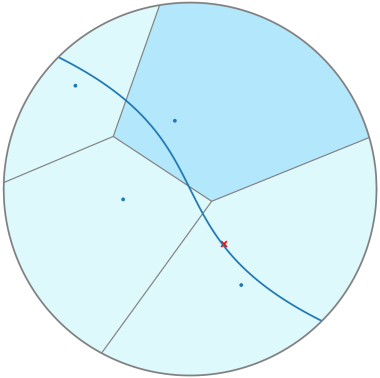
Worst-case sequence

Time	3
Mistake counter	3



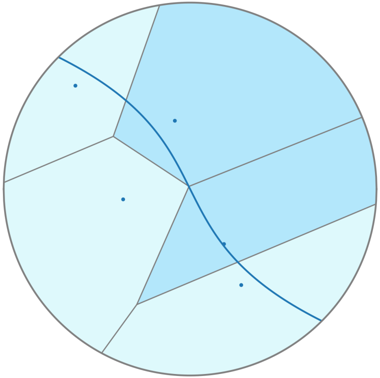
Worst-case sequence

Time	4
Mistake counter	4



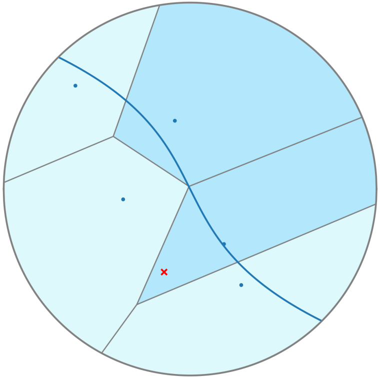
Worst-case sequence

Time	4
Mistake counter	4



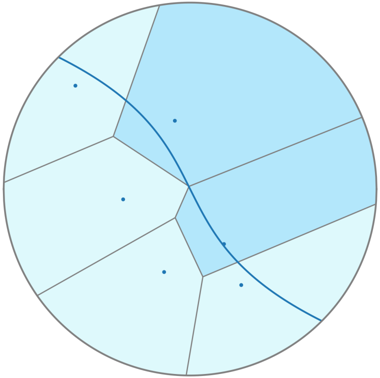
Worst-case sequence

Time	5
Mistake counter	5



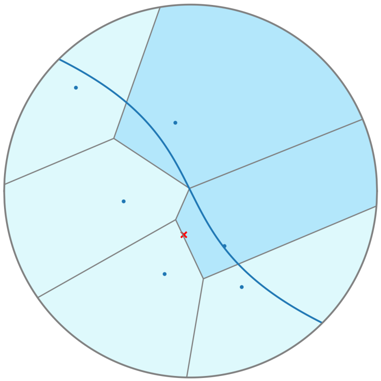
Worst-case sequence

Time	5
Mistake counter	5



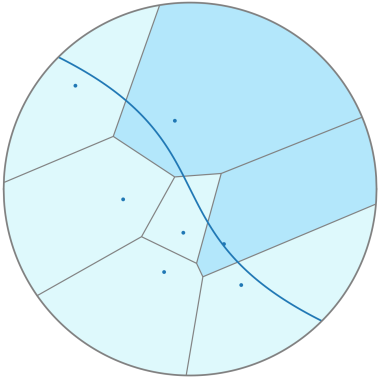
Worst-case sequence

Time	6
Mistake counter	6



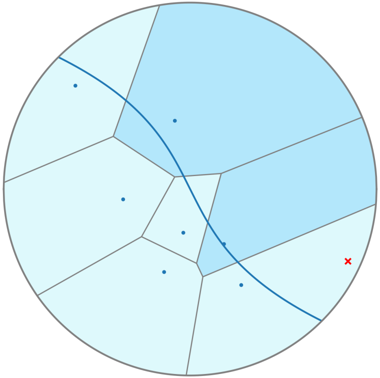
Worst-case sequence

Time	6
Mistake counter	6



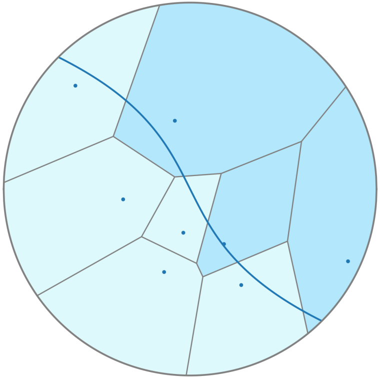
Worst-case sequence

Time	7
Mistake counter	7



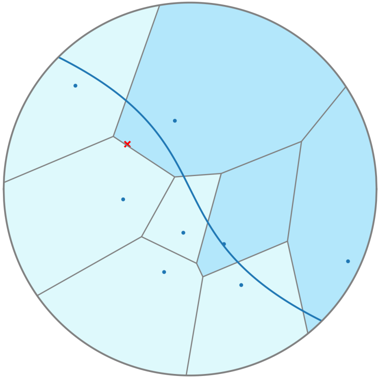
Worst-case sequence

Time	7
Mistake counter	7



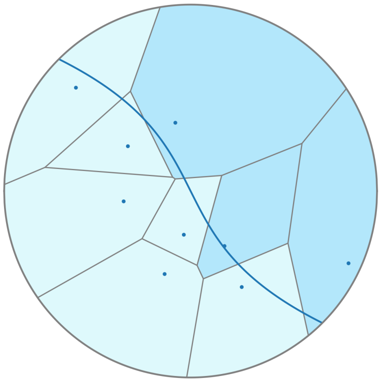
Worst-case sequence

Time	8
Mistake counter	8



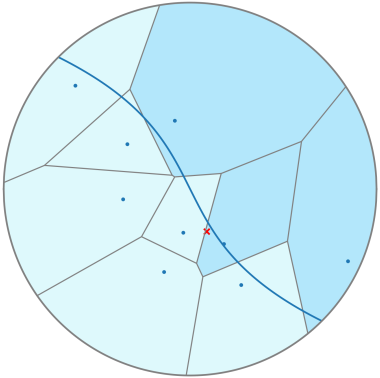
Worst-case sequence

Time	8
Mistake counter	8



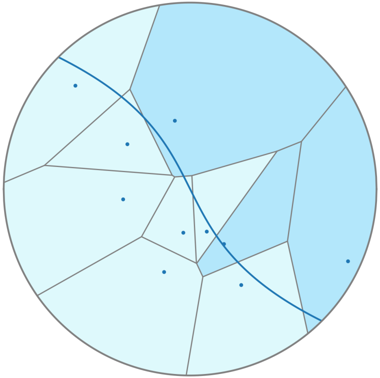
Worst-case sequence

Time	9
Mistake counter	9



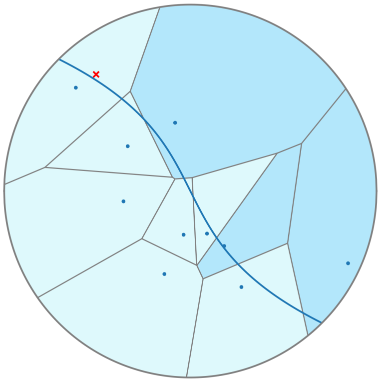
Worst-case sequence

Time	9
Mistake counter	9



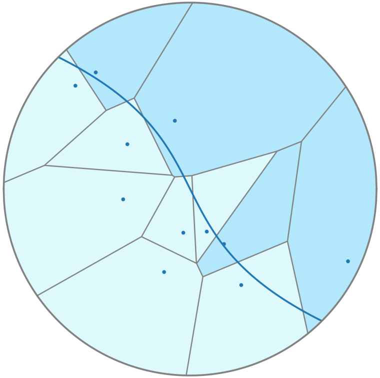
Worst-case sequence

Time	10
Mistake counter	10



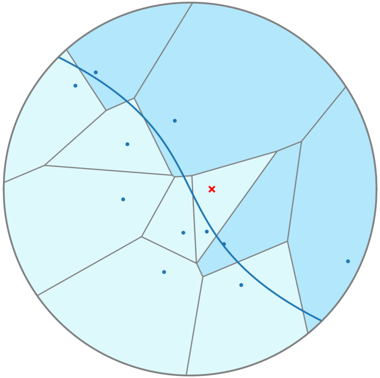
Worst-case sequence

Time	10
Mistake counter	10



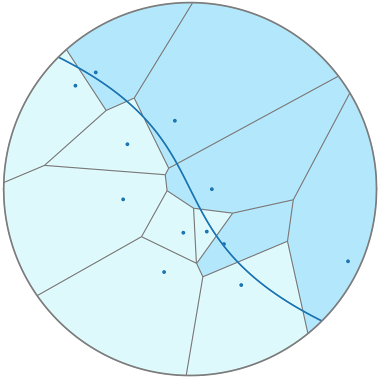
Worst-case sequence

Time	11
Mistake counter	11



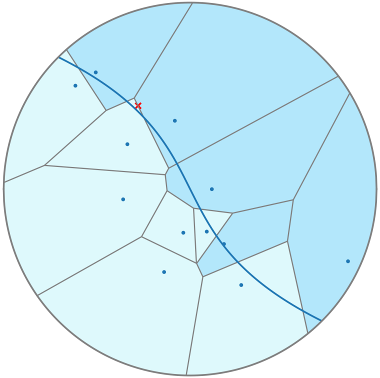
Worst-case sequence

Time	11
Mistake counter	11



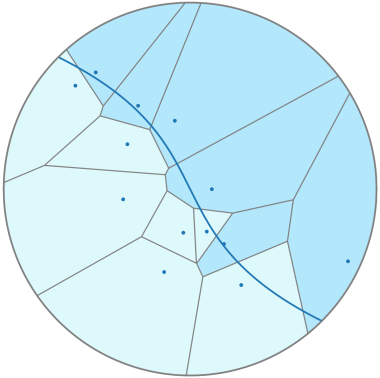
Worst-case sequence

Time	12
Mistake counter	12



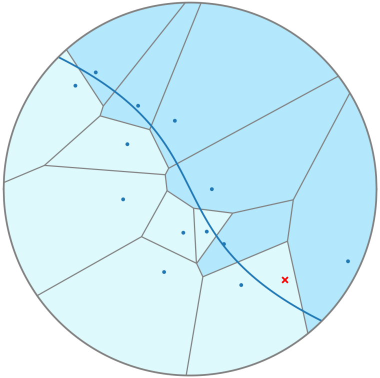
Worst-case sequence

Time	12
Mistake counter	12



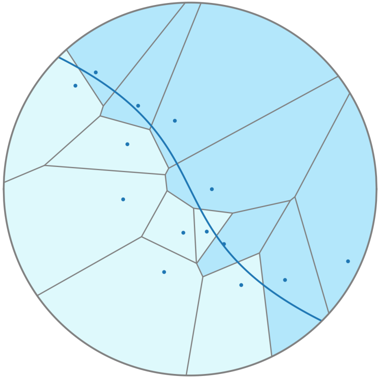
Worst-case sequence

Time	13
Mistake counter	13



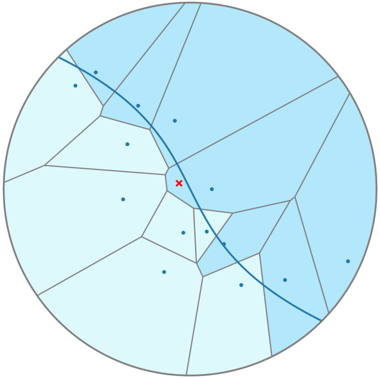
Worst-case sequence

Time	13
Mistake counter	13



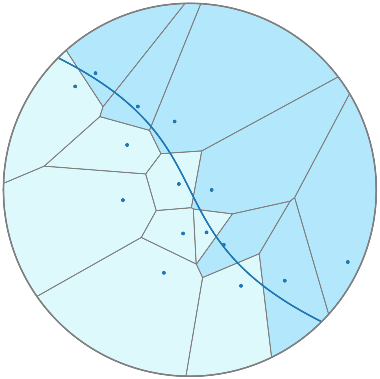
Worst-case sequence

Time	14
Mistake counter	14



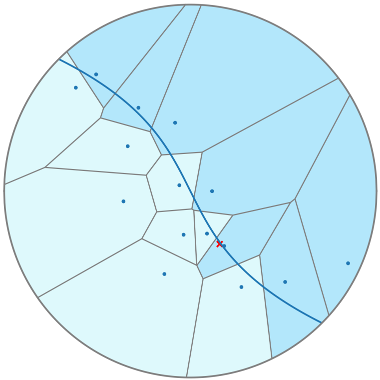
Worst-case sequence

Time	14
Mistake counter	14



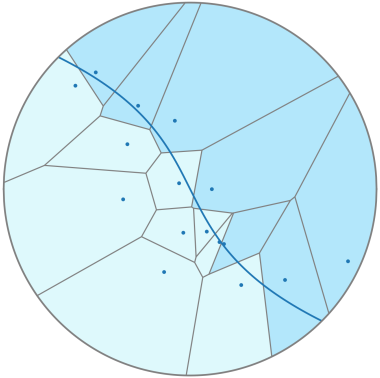
Worst-case sequence

Time	15
Mistake counter	15



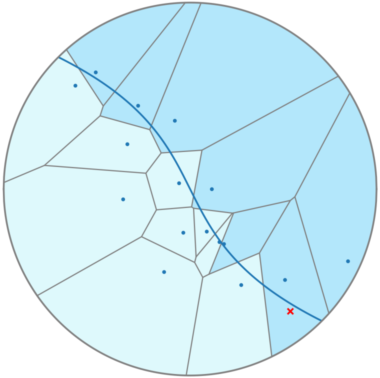
Worst-case sequence

Time	15
Mistake counter	15



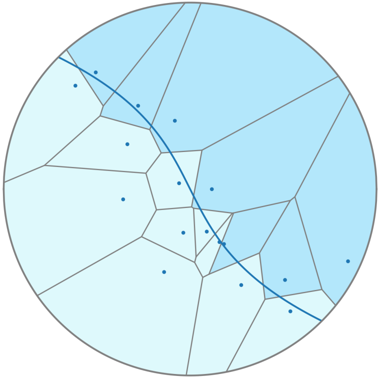
Worst-case sequence

Time	16
Mistake counter	16



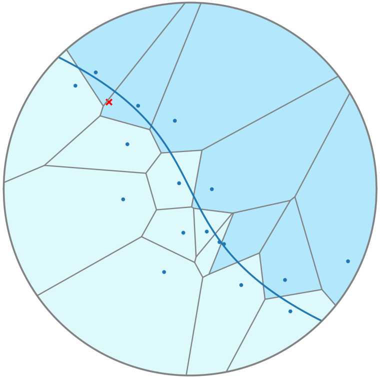
Worst-case sequence

Time	16
Mistake counter	16



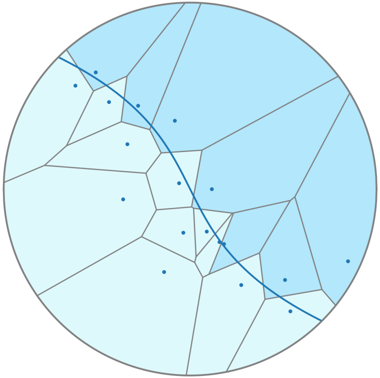
Worst-case sequence

Time	17
Mistake counter	17



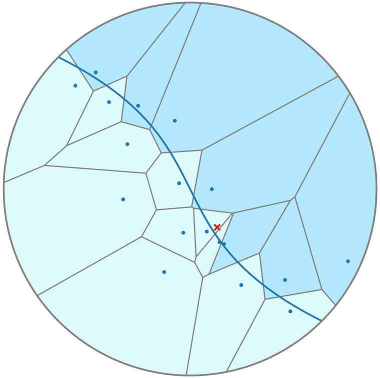
Worst-case sequence

Time	17
Mistake counter	17



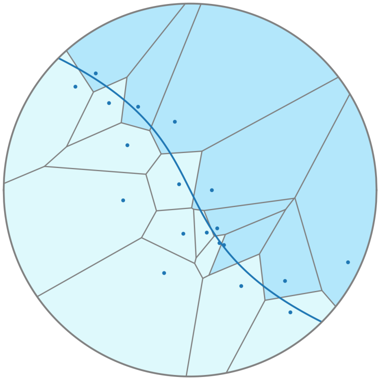
Worst-case sequence

Time	18
Mistake counter	18



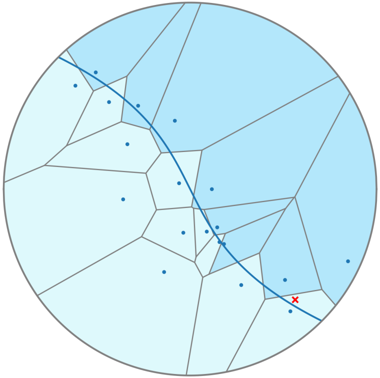
Worst-case sequence

Time	18
Mistake counter	18



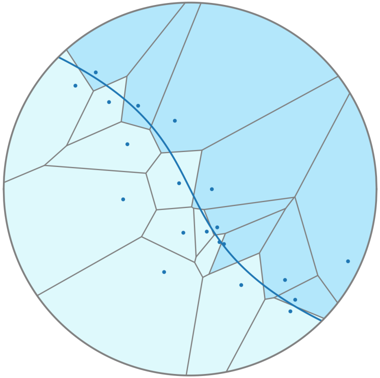
Worst-case sequence

Time	19
Mistake counter	19



Worst-case sequence

Time	19
Mistake counter	19



Question. When is the nearest neighbor rule consistent in the worst case?

Question. When is the nearest neighbor rule consistent in the worst case?

Answer. When different classes have positive separation.

A worst-case negative result

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space.

A worst-case negative result

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space.

Proposition

There exists a sequence $\mathbb{X}$ on which the nearest neighbor rule is not consistent on $(\mathbb{X}, \eta)$ if and only if the classes are not separated:

$$\inf_{\eta(x)\neq\eta(x')} \rho(x, x') = 0.$$

A worst-case negative result

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space.

Proposition

There exists a sequence $\mathbb{X}$ on which the nearest neighbor rule is not consistent on $(\mathbb{X}, \eta)$ if and only if the classes are not separated:

$$\inf_{\eta(x) \neq \eta(x')} \rho(x, x') = 0.$$

- The nearest neighbor version of having only finitely many things to learn.

Consistent settings for 1-nearest neighbor

i.i.d.

Cover and Hart (1967)

η has negligible boundary

$\mathcal{X}$ is separable

arbitrary

Kulkarni and Posner (1995)

η has separated classes

$\mathcal{X}$ is totally bounded

Question. How pathological are these worst-case sequences?

Question. How pathological are these worst-case sequences?

Answer. Extremely. Under mild conditions, they almost never occur.

Consistency for functions with negligible boundaries

Inductive bias of the nearest neighbor rule

Each point, once zoomed in enough, is surrounded by points of the same label.

Inductive bias of the nearest neighbor rule

Each point, once zoomed in enough, is surrounded by points of the same label.

This section.

Consistency when the inductive bias is correct almost everywhere.

↰ *for functions with negligible boundaries*

Metric measure space

Let $\mathcal{X}$ be a space with a separable metric ρ and a finite Borel measure ν.

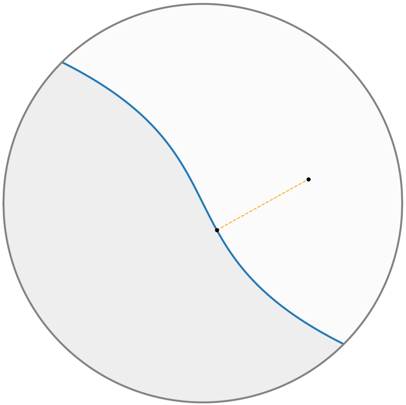
- Separable: every open cover has a countable subcover.
- Borel: we can measure the mass of balls.

Classification margin

Definition

The ***margin*** *of x with respect to η is given by:*

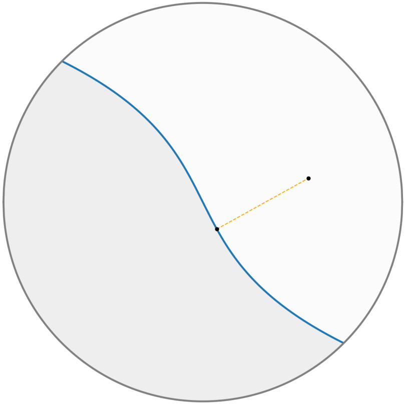
$$\mathrm{margin}_\eta(x) = \inf_{\eta(x)\neq\eta(x')} \rho(x, x').$$



Functions with negligible boundaries

Definition

A function η has **negligible boundary** *if ν-almost all points have positive margin.*



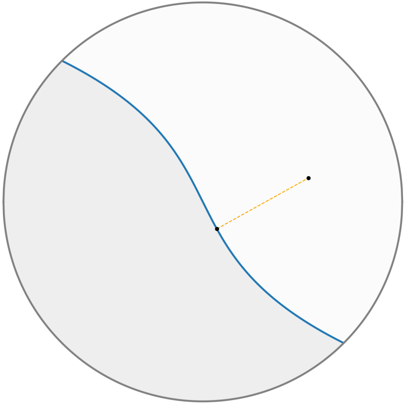
Functions with negligible boundaries

Definition

A function η has **negligible boundary** *if ν-almost all points have positive margin.*

Example

Let $\mathcal{X}$ be Euclidean space with the Lebesgue measure. Let η have smooth decision boundary.

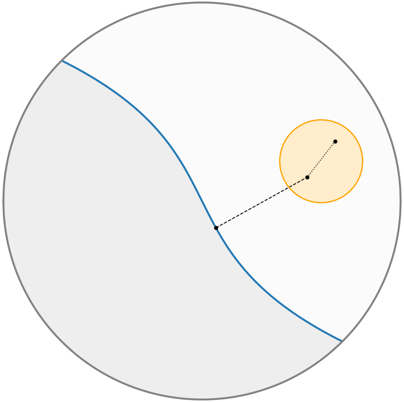


Mutually-labeling set

Definition

A set $U \subset \mathcal{X}$ is mutually-labeling for η when:

$$\mathrm{diam}(U) < \mathrm{margin}_\eta(x), \qquad \forall x \in U.$$



Mutually-labeling set

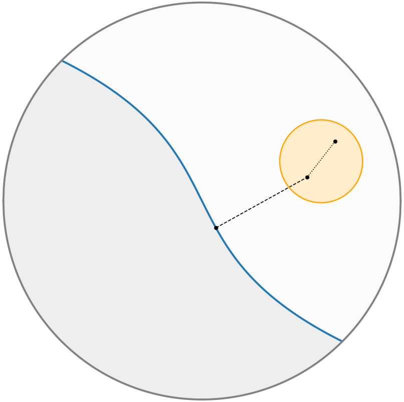
Definition

*A set $U \subset \mathcal{X}$ is **mutually-labeling** for η when:*

$$\operatorname{diam}(U) < \operatorname{margin}_\eta(x), \qquad \forall x \in U.$$

Proposition

*For all time, the nearest neighbor rule makes at most **one mistake per mutually-labeling set**.*



Mutually-labeling set

Definition

*A set $U \subset \mathcal{X}$ is **mutually-labeling** for η when:*

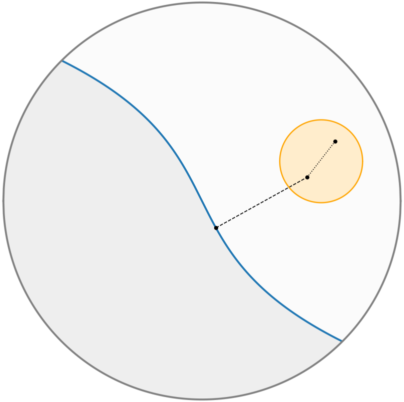
$$\operatorname{diam}(U) < \operatorname{margin}_\eta(x), \qquad \forall x \in U.$$

Proposition

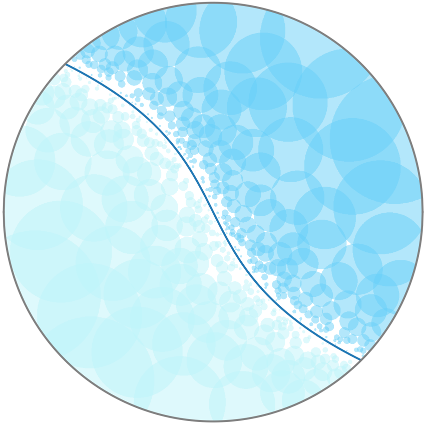
Let x have positive margin:

$$r_x = \operatorname{margin}_\eta(x) > 0.$$

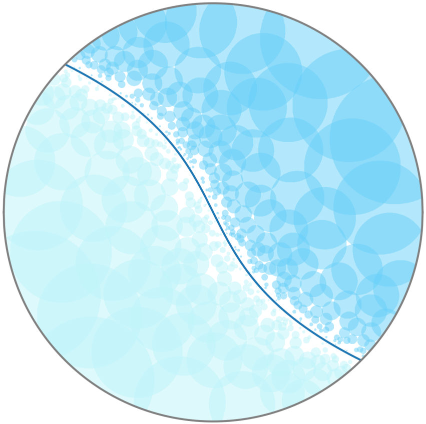
The open ball $B(x, r_x/3)$ is mutually labeling.



Mutually-labeling cover

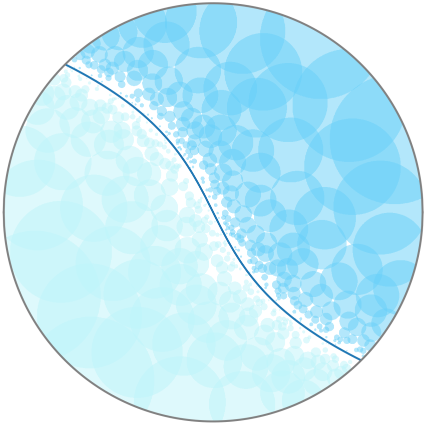


Mutually-labeling cover



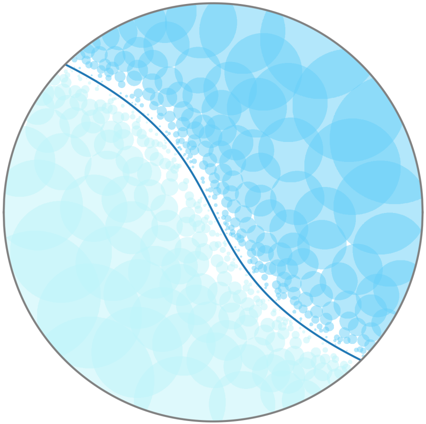
1. η has negligible boundary $\Longrightarrow$
   mutually-labeling ball cover for $\mathcal{X}$ a.e.

Mutually-labeling cover



1. η has negligible boundary $\implies$
   mutually-labeling ball cover for $\mathcal{X}$ a.e.
2. ρ is a separable metric $\implies$
   countable subcover

Mutually-labeling cover



1. η has negligible boundary $\Longrightarrow$
   mutually-labeling ball cover for $\mathcal{X}$ a.e.
2. ρ is a separable metric $\Longrightarrow$
   countable subcover
3. ν is a finite measure $\Longrightarrow$
   finite, arbitrarily-good approximate cover

Let η have negligible boundary.

Let η have negligible boundary. Eventually, all mistakes made by the nearest neighbor rule must come from an arbitrarily small region w.r.t. ν.

Let η have negligible boundary. Eventually, all mistakes made by the nearest neighbor rule must come from an arbitrarily small region w.r.t. ν.

What is the rate that $\mathbb{X}$ lands in regions with arbitrarily small mass?

Let η have negligible boundary. Eventually, all mistakes made by the nearest neighbor rule must come from an arbitrarily small region w.r.t. ν.

What is the rate that $\mathbb{X}$ lands in regions with arbitrarily small mass?

↳ *if this rate goes to zero, then the nearest neighbor rule is consistent*

Stochastic processes with a time-averaged constraint

Definition (Ergodic continuity)

A stochastic process $\mathbb{X}$ *is* ***ergodically dominated*** *by* ν *if for all* $\varepsilon > 0$*, there is a* $\delta > 0$ *where:*

$$\nu(A) < \delta \quad \implies \quad \limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} 1\{X_n \in A\} < \varepsilon \quad \text{a.s.}$$

We say that $\mathbb{X}$ *is* ***ergodically continuous*** *with respect to* ν *at rate* $\varepsilon(\delta)$*.*

Stochastic processes with a time-averaged constraint

Definition (Ergodic continuity)

A stochastic process $\mathbb{X}$ is ergodically dominated by ν if for all $\varepsilon > 0$, there is a $\delta > 0$ where:

$$\nu(A) < \delta \quad \implies \quad \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A\} < \varepsilon \quad \text{a.s.}$$

We say that $\mathbb{X}$ is ergodically continuous with respect to ν at rate $\varepsilon(\delta)$.

Interpretations.

- $\mathbb{X}$ comes from a *budgeted adversary.*
- The constraint is only on the *tail* of $\mathbb{X}$.
- The empirical submeasure $A \mapsto \limsup_{N\to\infty} \frac{1}{N} \sum \mathbb{1}\{X_n \in A\}$ is absolutely continuous with respect to ν.

Example of ergodic continuity

Definition (Ergodic continuity)

A stochastic process $\mathbb{X}$ is ergodically dominated by ν if for all $\varepsilon > 0$, there is a $\delta > 0$ where:

$$\nu(A) < \delta \quad \Longrightarrow \quad \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A\} < \varepsilon \quad \text{a.s.}$$

We say that $\mathbb{X}$ is ergodically continuous with respect to ν at rate $\varepsilon(\delta)$.

I.I.D. processes are ergodically dominated.

- Apply the law of large numbers.

Consistency for nice functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be a space where ρ is a separable metric and ν is a finite Borel measure.

Consistency for nice functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be a space where ρ is a separable metric and ν is a finite Borel measure. Suppose that $\mathbb{X}$ is ergodically dominated by ν and η has negligible boundary.

Consistency for nice functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be a space where ρ is a separable metric and ν is a finite Borel measure. Suppose that $\mathbb{X}$ is ergodically dominated by ν and η has negligible boundary. Then:

$$\underbrace{\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\eta(X_n) \neq \eta(\tilde{X}_n)\} = 0}_{\text{the nearest neighbor rule is online consistent for } (\mathbb{X}, \eta).} \qquad \text{a.s.}$$

Consistent settings for 1-nearest neighbor

η has negligible boundary
$\mathcal{X}$ is separable

ergodically dominated

i.i.d.

arbitrary

Cover and Hart (1967)
η has negligible boundary
$\mathcal{X}$ is separable

Kulkarni and Posner (1995)
η has separated classes
$\mathcal{X}$ is totally bounded

Universal consistency on upper doubling spaces

Universal consistency

Goal: consistency for all measurable functions almost surely.

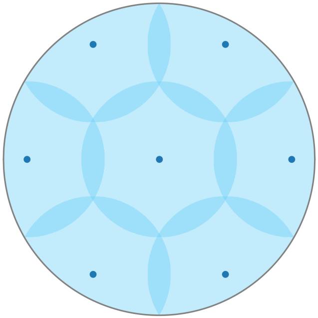
Universal consistency

Goal: consistency for all measurable functions almost surely.

- Boundary points are no longer localized to a measure zero set.
  - e.g. $\eta(x) = \mathbb{1}\{x \in \mathbb{Q}\}$.

Introducing a geometric assumption

Definition
A metric space $(\mathcal{X}, \rho, \nu)$ is **doubling** *when each ball can be covered by at most 2^d balls of half its radius.*



Approximation by functions with negligible boundary

Let ρ be a doubling metric and ν a finite Borel measure.

Approximation by functions with negligible boundary

Let ρ be a doubling metric and ν a finite Borel measure.

Proposition

The set of functions with negligible boundary is dense in $L^1(\mathcal{X}; \nu)$.

Approximation by functions with negligible boundary

Let ρ be a doubling metric and ν a finite Borel measure.

Proposition

The set of functions with negligible boundary is dense in $L^1(\mathcal{X};\nu)$.

↰ Key ingredient: a Lebesgue differentiation theorem on doubling spaces.

A reasonable conjecture.
Approximate η very well by some η' with negligible boundary.

A reasonable conjecture.
Approximate η very well by some η' with negligible boundary.

- When $\mathbb{X}$ is ergodically dominated, learning η is like learning η' when they have vanishingly small disagreement region $\{\eta \neq \eta'\}$.

A reasonable conjecture.
Approximate η very well by some η' with negligible boundary.

- When $\mathbb{X}$ is ergodically dominated, learning η is like learning η' when they have vanishingly small disagreement region $\{\eta \neq \eta'\}$.

  ↰ Since X_n rarely lands in $\{\eta \neq \eta'\}$.

A reasonable conjecture.
Approximate η very well by some η' with negligible boundary.

- When $\mathbb{X}$ is ergodically dominated, learning η is like learning η' when they have vanishingly small disagreement region $\{\eta \neq \eta'\}$.

  ↰ Since X_n rarely lands in $\{\eta \neq \eta'\}$.

This turns out to be wrong.

- Blanchard (2022) constructs example where 1-NN is not consistent, but $\mathcal{X} = [0, 1]$ is 1-doubling, η is measurable, and $\mathbb{X}$ is ergodically dominated.

What goes wrong?

- The influence of $\{\eta \neq \eta'\}$ is not limited to the times that X_n lands in it.

What goes wrong?

- The influence of $\{\eta \neq \eta'\}$ is not limited to the times that X_n lands in it.
- Those instances can be the nearest neighbor of downstream points.

What goes wrong?

- The influence of $\{\eta \neq \eta'\}$ is not limited to the times that X_n lands in it.
- Those instances can be the nearest neighbor of downstream points.

Insufficiency of a tail constraint.
'Bad points' can accumulate in memory, and their influence grows and shrinks with their Voronoi cells.

What goes wrong?

- The influence of $\{\eta \neq \eta'\}$ is not limited to the times that X_n lands in it.
- Those instances can be the nearest neighbor of downstream points.

Insufficiency of a tail constraint.
'Bad points' can accumulate in memory, and their influence grows and shrinks with their Voronoi cells.

- A new problem: the 'hard part' changes over time.

Stochastic processes with a time-uniform constraint

Definition (Uniform absolute continuity)

A stochastic process $\mathbb{X}$ is uniformly dominated by ν if for all $\varepsilon > 0$, there is a $\delta > 0$ where:

$$\nu(A) < \delta \quad \implies \quad \Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right) < \varepsilon \quad \text{a.s.}$$

We say that $\mathbb{X}$ is uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$.

Stochastic processes with a time-uniform constraint

Definition (Uniform absolute continuity)

A stochastic process $\mathbb{X}$ is uniformly dominated by ν if for all $\varepsilon > 0$, there is a $\delta > 0$ where:

$$\nu(A) < \delta \quad \Longrightarrow \quad \Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right) < \varepsilon \quad \text{a.s.}$$

We say that $\mathbb{X}$ is uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$.

Interpretations.

- $\mathbb{X}$ comes from a *bounded precision adversary.*
- The constraint is strictly stronger, and applies to each point in time.
- Ergodic continuity is retrospective; this is a generative constraint.

Ergodic continuity v. uniform absolute continuity

- Ergodic continuity: looking back, how often did points land in A?

Ergodic continuity v. uniform absolute continuity

- **Ergodic continuity:** looking back, how often did points land in A?
- **Uniform absolute continuity:** how easily can an adversary generate a point from A?

Ergodic continuity v. uniform absolute continuity

- Ergodic continuity: looking back, how often did points land in A?
  ↳ *helpful when hard regions are fixed in space*
- Uniform absolute continuity: how easily can an adversary generate a point from A?

Ergodic continuity v. uniform absolute continuity

- Ergodic continuity: looking back, how often did points land in A?
  ↳ *helpful when hard regions are fixed in space*
- Uniform absolute continuity: how easily can an adversary generate a point from A?
  ↳ *helpful when hard regions change over time*

Ergodic continuity v. uniform absolute continuity

- Ergodic continuity: looking back, how often did points land in A?
  ↰ *helpful when hard regions are fixed in space*
- Uniform absolute continuity: how easily can an adversary generate a point from A?
  ↰ *helpful when hard regions change over time*

Uniformly dominated processes are ergodically dominated.

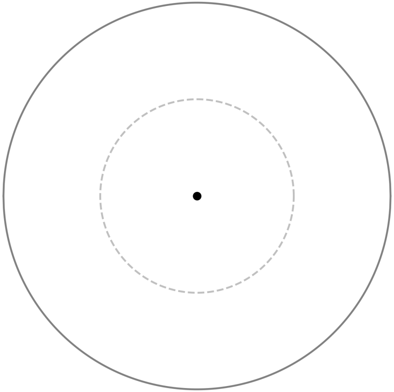
- Apply the martingale law of large numbers.

Why is uniform absolute continuity helpful?

- **A simpler problem:** bound the influence of a single point X_0 on $\tilde{\mathbb{X}}$.

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{1}\{\tilde{X}_n = X_0\}\right]$$

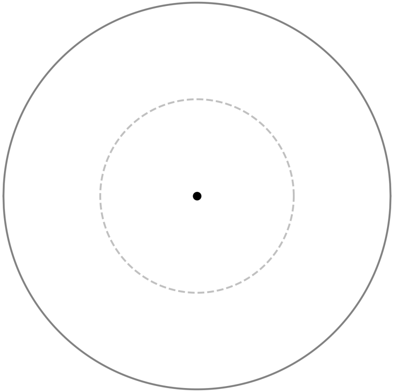
Influence of a single point



Metric entropy bound

How many times can the following occur?

Influence of a single point

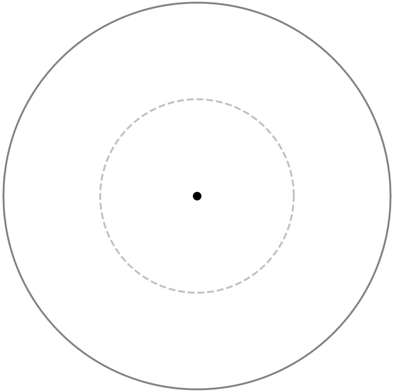


Metric entropy bound

How many times can the following occur?

- X_0 is a nearest neighbor of X_n

Influence of a single point

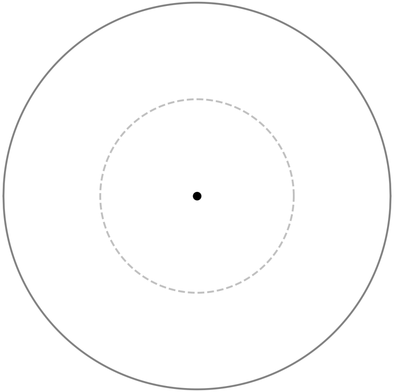


Metric entropy bound

How many times can the following occur?

- X_0 is a nearest neighbor of X_n
- They are r-separated $\rho(X_0, X_n) > r$

Influence of a single point



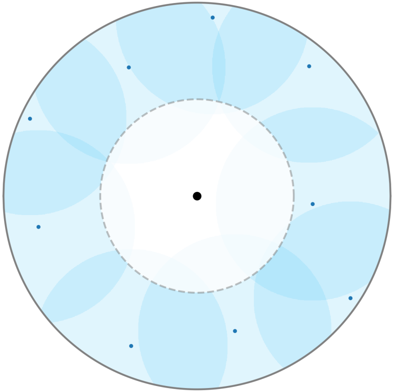
Metric entropy bound

How many times can the following occur?

- X_0 is a nearest neighbor of X_n
- They are r-separated $\rho(X_0, X_n) > r$

Answer: the r-packing number of the space.

Influence of a single point

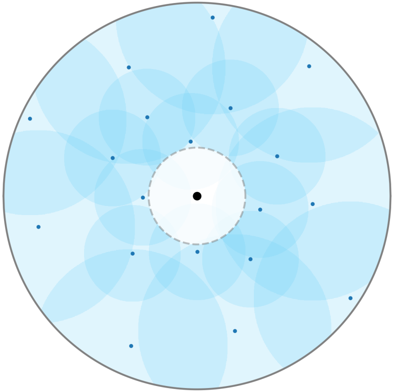


In a doubling space with unit diameter

X_0 is a nearest neighbor of:

- points in $B(X_0, 1/2)^c$ at most 2^d times

Influence of a single point

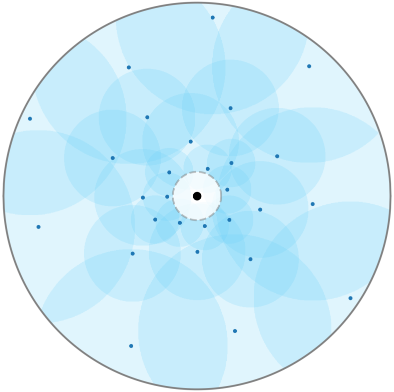


In a doubling space with unit diameter

X_0 is a nearest neighbor of:

- points in $B(X_0, 1/4)^c$ at most $2 \cdot 2^d$ times

Influence of a single point

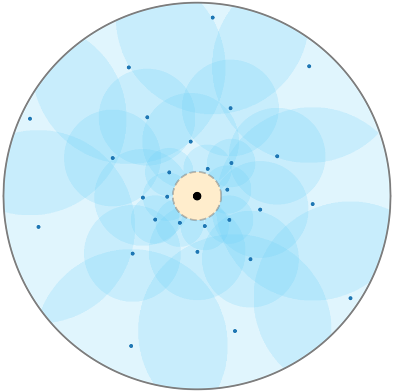


In a doubling space with unit diameter

X_0 is a nearest neighbor of:

- points in $B(X_0, 1/2^k)^c$ at most $k \cdot 2^d$ times

Influence of a single point



In a doubling space with unit diameter

X_0 is a nearest neighbor of:

- points in $B(X_0, 1/2^k)^c$ at most $k \cdot 2^d$ times
- points in $B(X_0, 1/2^k)$ with small probability

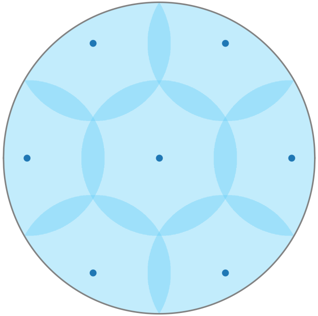
  by uniform absolute continuity ⤴

Upper doubling measure

Definition

A d-doubling space has an upper doubling measure if:

$$\nu\big(B(x, r)\big) \leq c r^d.$$



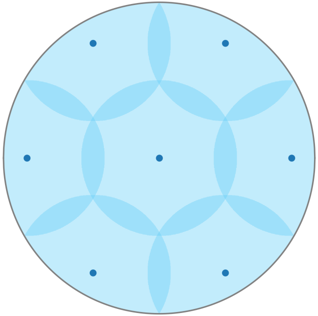
Upper doubling measure

Definition

A d-doubling space has an **upper doubling** *measure if:*

$$\nu\big(B(x,r)\big) \leq cr^d.$$

Then, a set with small metric entropy has small measure.



What is the influence of a single point on $\tilde{\mathbb{X}}$?

- If $(\mathcal{X}, \rho, \nu)$ is upper doubling and $\mathbb{X}$ is uniformly dominated at rate $\varepsilon(\delta)$,

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\}\right] \leq \frac{k \cdot 2^d}{N} + \varepsilon\left(c2^{-k}\right), \qquad \forall k \in \mathbb{N}.$$

Idea for universal consistency

1. Even though 'bad points' can accumulate in memory, in a doubling space, their Voronoi cells tend to quickly shrink (in the metric entropy sense) as they are hit.

Idea for universal consistency

1. Even though 'bad points' can accumulate in memory, in a doubling space, their Voronoi cells tend to quickly shrink (in the metric entropy sense) as they are hit.
2. These Voronoi cells also shrink with respect to ν in upper doubling spaces.

Idea for universal consistency

1. Even though 'bad points' can accumulate in memory, in a doubling space, their Voronoi cells tend to quickly shrink (in the metric entropy sense) as they are hit.
2. These Voronoi cells also shrink with respect to ν in upper doubling spaces.
3. Then, it becomes increasingly unlikely that these bad points are nearest neighbors if $\mathbb{X}$ is uniformly dominated.

Ergodic continuity of the nearest neighbor process

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be bounded and upper doubling.

Ergodic continuity of the nearest neighbor process

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be bounded and upper doubling. Let $\mathbb{X}$ be uniformly dominated at rate $\varepsilon(\delta)$.

Ergodic continuity of the nearest neighbor process

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be bounded and upper doubling. Let $\mathbb{X}$ be uniformly dominated at rate $\varepsilon(\delta)$. Then, the nearest neighbor process $\tilde{\mathbb{X}}$ is ergodically dominated at rate $O(\varepsilon(\delta) \log \frac{1}{\delta})$.

Ergodic continuity of the nearest neighbor process

Theorem
Let $(\mathcal{X}, \rho, \nu)$ be bounded and upper doubling. Let $\mathbb{X}$ be uniformly dominated at rate $\varepsilon(\delta)$. Then, the nearest neighbor process $\tilde{\mathbb{X}}$ is ergodically dominated at rate $O(\varepsilon(\delta) \log \frac{1}{\delta})$.

In words:
Let η and η' rarely disagree. The average rate that $\tilde{\mathbb{X}}$ lands in $\{\eta \neq \eta'\}$ is tiny.

Consistency for all measurable functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ *be upper doubling,*

Consistency for all measurable functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be upper doubling, where ρ is separable and ν is finite. Let η be measurable. Suppose that $\mathbb{X}$ is uniformly dominated by ν.

Consistency for all measurable functions

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be upper doubling, where ρ is separable and ν is finite. Let η be measurable. Suppose that $\mathbb{X}$ is uniformly dominated by ν. Then:

$$\underbrace{\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{\eta(X_n) \neq \eta(\tilde{X}_n)\} = 0}_{\text{the nearest neighbor rule is online consistent for } (\mathbb{X}, \eta).} \qquad \text{a.s.}$$

Proof sketch

1. Let η be approximated arbitrarily well by η' with negligible boundary.

Proof sketch

1. Let η be approximated arbitrarily well by η' with negligible boundary.
2. $\mathbb{X}$ is uniformly dominated, so the mistake rate on η' vanishes.

Proof sketch

1. Let η be approximated arbitrarily well by η' with negligible boundary.
2. $\mathbb{X}$ is uniformly dominated, so the mistake rate on η' vanishes.
3. If the mistake rate on η does not vanish, this must be due to $\{\eta \neq \eta'\}$.

Proof sketch

1. Let η be approximated arbitrarily well by η' with negligible boundary.
2. $\mathbb{X}$ is uniformly dominated, so the mistake rate on η' vanishes.
3. If the mistake rate on η does not vanish, this must be due to $\{\eta \neq \eta'\}$.
4. But the nearest neighbor process cannot significantly amplify influence of arbitrarily small regions, implying universal consistency.

Consistency of the nearest neighbor rule

Takeaways and open problems

Non-worst-case online learning

Motif of smoothed analysis

While worst-case analyses provide important safeguards, they can be too pessimistic.

- They can fail to explain observed behavior.

Non-worst-case online learning

Motif of smoothed analysis

While worst-case analyses provide important safeguards, they can be too pessimistic.

- They can fail to explain observed behavior.
- What constitutes a ‘typical’ online sequence of tasks?

Constrained classes of stochastic processes

i.i.d. $\subset$ smoothed $\subset$ uniformly dominated $\subset$ ergodically dominated $\subset$ $\mathcal{C}_1$ $\subset$ arbitrary

- **Smoothed processes**:
  (Rakhlin et al., 2011; Haghtalab et al., 2020, 2022; Block et al., 2022)
- **Online learnable processes**:
  (Hanneke, 2021; Blanchard and Cosson, 2022; Blanchard, 2022)

Open problems

1. **Benign noise:** when does the k_n-nearest neighbor rule learn?
2. **Bounded memory:** when is bounded memory sufficient?
3. **Adaptive rates:** can we get meaningful/problem-dependent rates?

Thank you!

Sanjoy Dasgupta and Geelon So. Online Consistency of the Nearest Neighbor Rule. In *The 38th Conference on Neural Information Processing Systems*, 2024.

References I

Moise Blanchard. Universal online learning: An optimistically universal learning rule. In *Conference on Learning Theory*, pages 1077–1125. PMLR, 2022.

Moise Blanchard and Romain Cosson. Universal online learning with bounded loss: Reduction to binary classification. In *Conference on Learning Theory*, pages 479–495. PMLR, 2022.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, pages 1716–1786. PMLR, 2022.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis, nonparametric discrimination. *USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004, Report 4, Contract AD41(128)-31*, 1951.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 33:9203–9215, 2020.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 942–953. IEEE, 2022.

References II

Steve Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22(130):1–116, 2021.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic and constrained adversaries. *arXiv preprint arXiv:1104.5070*, 2011.