Learning when learning is reasonable

Thesis Proposal

Geelon So, geelon@ucsd.edu
Advisors: Sanjoy Dasgupta and Yian Ma
June 5, 2024

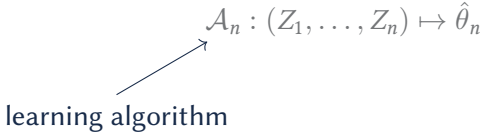
Outline of talk

1. Motivation
2. Online classification
3. Non-worst case sequences
4. Consistency of the nearest neighbor rule
5. Ergodic continuity of nearest neighbor processes
6. Looking forward

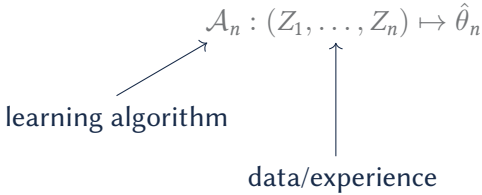
A working definition of learning

$$\mathcal{A}_n : (Z_1, \ldots, Z_n) \mapsto \hat{\theta}_n$$

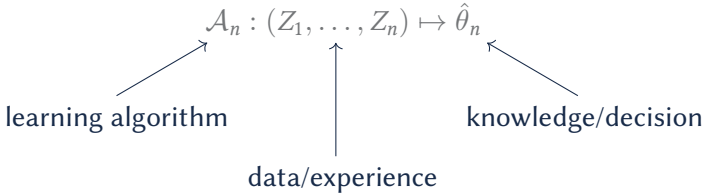
A working definition of learning



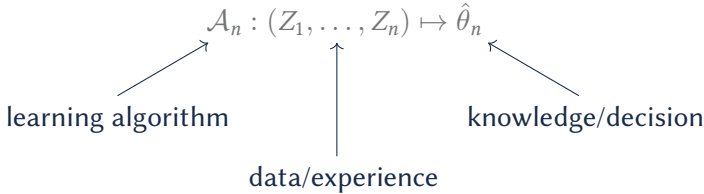
A working definition of learning



A working definition of learning



A working definition of learning



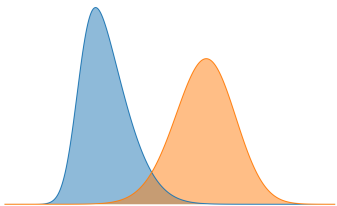
Learning: to derive better knowledge/decisions with more data/experience.

Question. What do past experiences have to do with the current task?

Question. What do past experiences have to do with the current task?

- Different assumptions about this lead to different models of learning.

Two standards of machine learning



Statistical (i.i.d.)

Online (worst-case)

Learning beyond I.I.D. and worst-case settings

Question. What are general yet tractable conditions for learning?

Learning beyond I.I.D. and worst-case settings

Question. What are general yet tractable conditions for learning?

This talk

Dasgupta, S. and So, G. *Online consistency of the nearest neighbor rule.* Submitted (2024).

Learning beyond I.I.D. and worst-case settings

Question. What are general yet tractable conditions for learning?

This talk

Dasgupta, S. and So, G. *Online consistency of the nearest neighbor rule.* Submitted (2024).

Contributions

- We introduce general conditions for learning that admit simple analyses.

Learning beyond I.I.D. and worst-case settings

Question. What are general yet tractable conditions for learning?

This talk

Dasgupta, S. and So, G. *Online consistency of the nearest neighbor rule.* Submitted (2024).

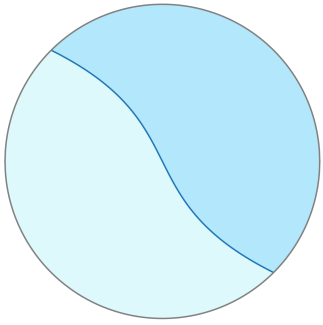
Contributions

- We introduce general conditions for learning that admit simple analyses.
- We show that the nearest neighbor rule is consistent under much broader conditions than previously known.

Online classification

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.

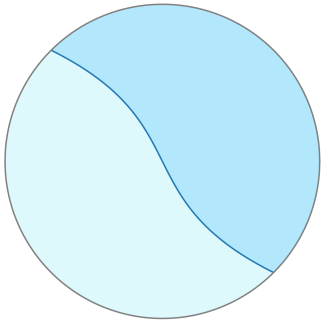


The realizable online setting

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.

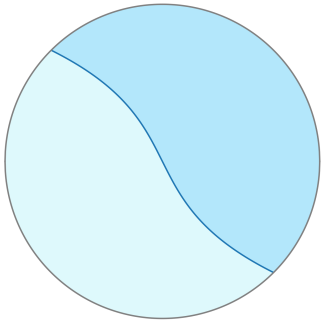


The realizable online setting

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 $\hat{Y}_n$.

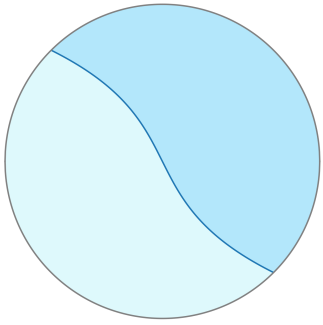


The realizable online setting

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.

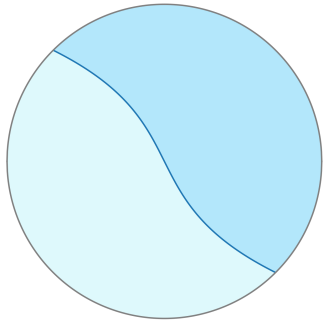
Online classification loop.
For $n = 1, 2, \ldots$

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



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

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

Nearest neighbor process

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

Definition

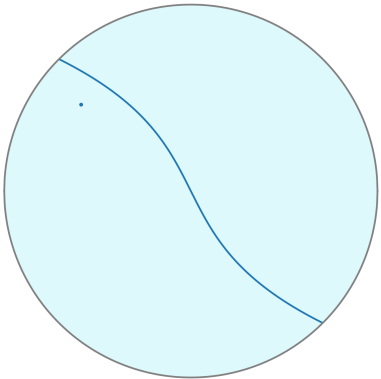
A nearest neighbor process is a sequence $\tilde{\mathbb{X}} = (\tilde{X}_n)_{n>0}$ satisfying

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

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

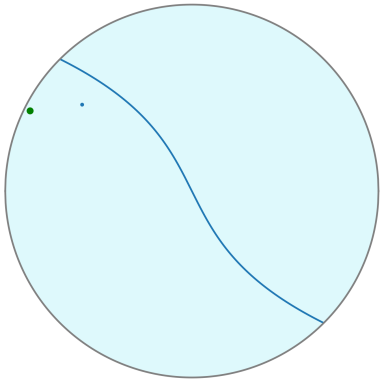
Behavior of the nearest neighbor rule in the i.i.d. setting.

I.I.D. sequence



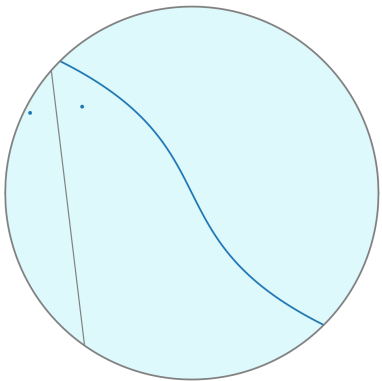
Time	0
Mistake counter	0

I.I.D. sequence



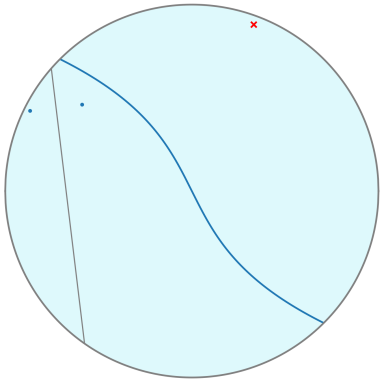
Time	1
Mistake counter	0

I.I.D. sequence



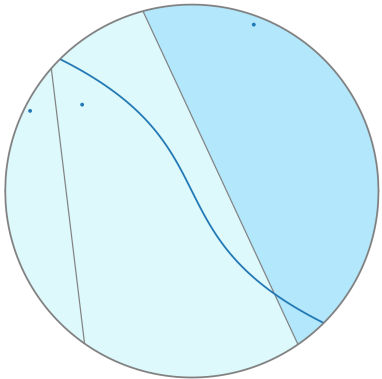
Time	1
Mistake counter	0

I.I.D. sequence



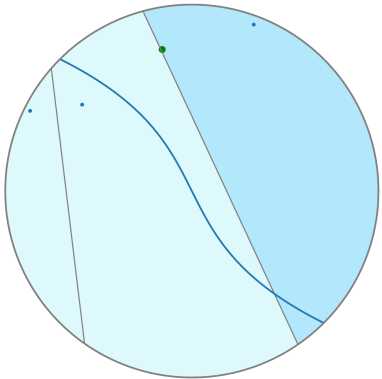
Time	2
Mistake counter	1

I.I.D. sequence



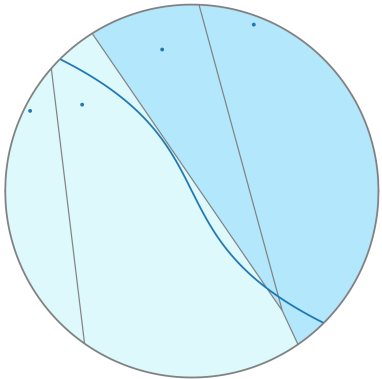
Time	2
Mistake counter	1

I.I.D. sequence



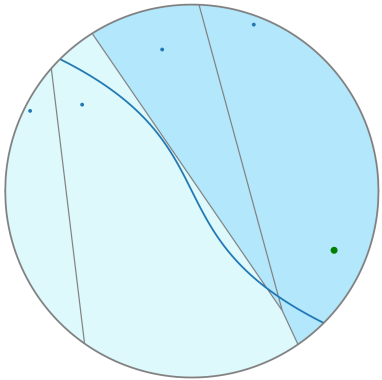
Time	3
Mistake counter	1

I.I.D. sequence



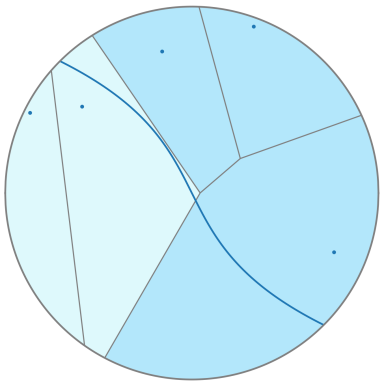
Time	3
Mistake counter	1

I.I.D. sequence



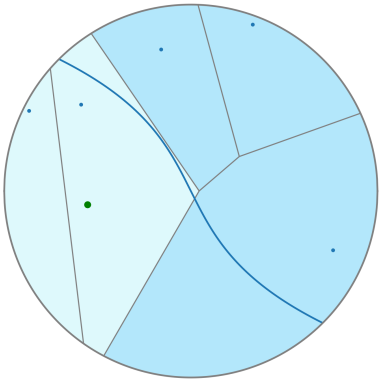
Time	4
Mistake counter	1

I.I.D. sequence



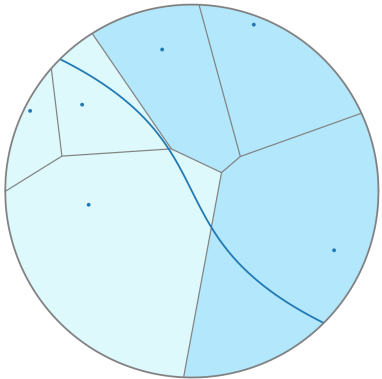
Time	4
Mistake counter	1

I.I.D. sequence



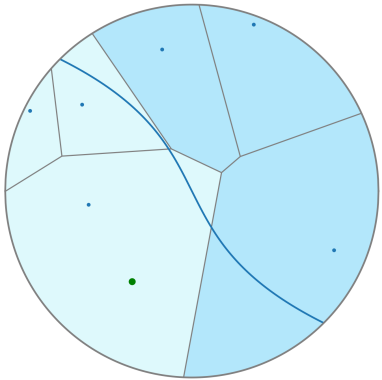
Time	5
Mistake counter	1

I.I.D. sequence



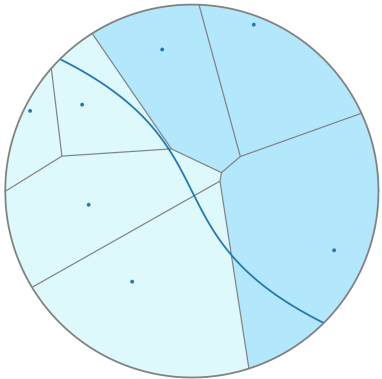
Time	5
Mistake counter	1

I.I.D. sequence



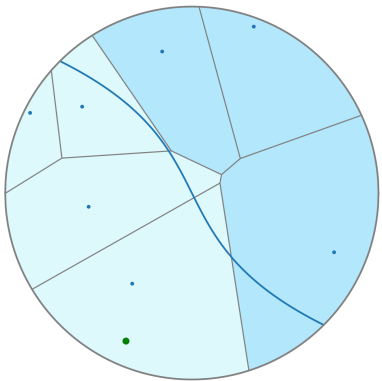
Time	6
Mistake counter	1

I.I.D. sequence



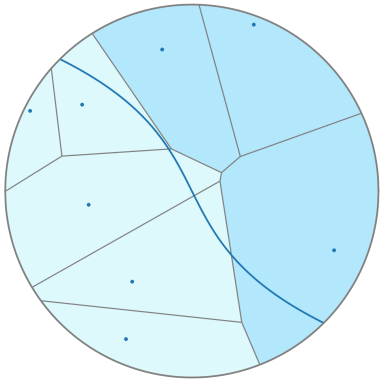
Time	6
Mistake counter	1

I.I.D. sequence



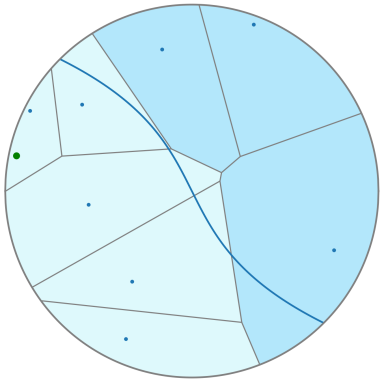
Time	7
Mistake counter	1

I.I.D. sequence



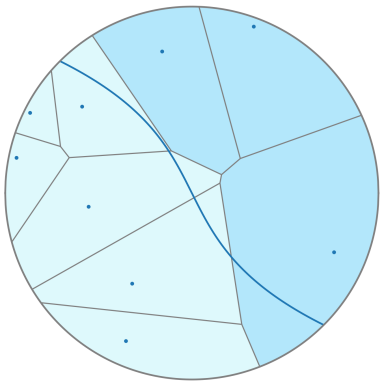
Time	7
Mistake counter	1

I.I.D. sequence



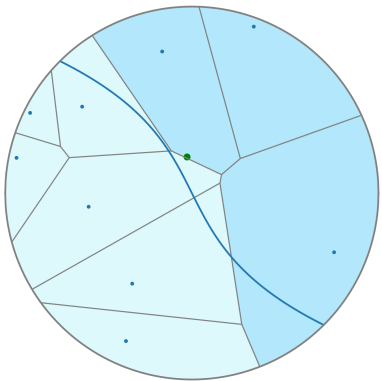
Time	8
Mistake counter	1

I.I.D. sequence



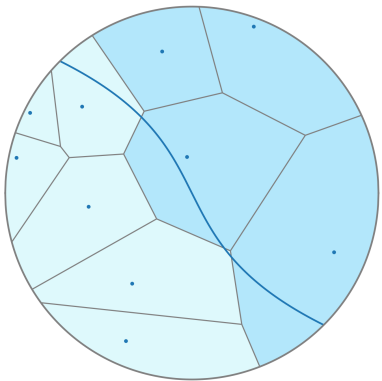
Time	8
Mistake counter	1

I.I.D. sequence



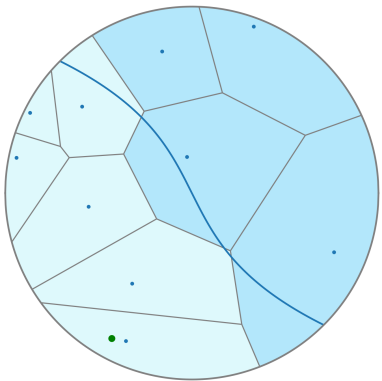
Time	9
Mistake counter	1

I.I.D. sequence



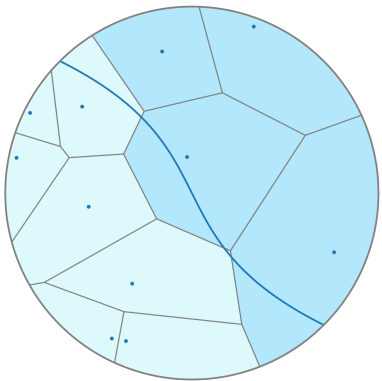
Time	9
Mistake counter	1

I.I.D. sequence



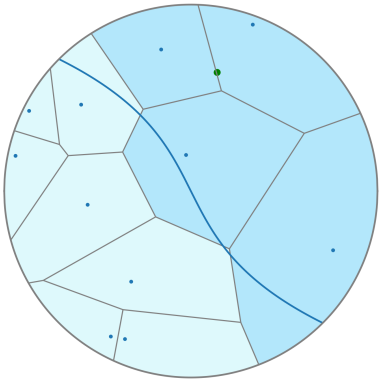
Time	10
Mistake counter	1

I.I.D. sequence



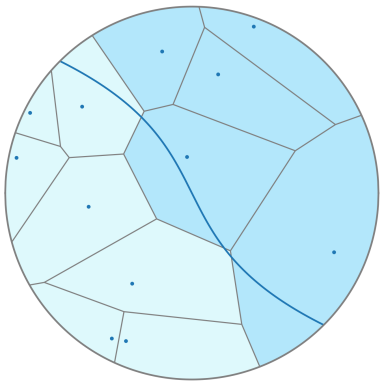
Time	10
Mistake counter	1

I.I.D. sequence



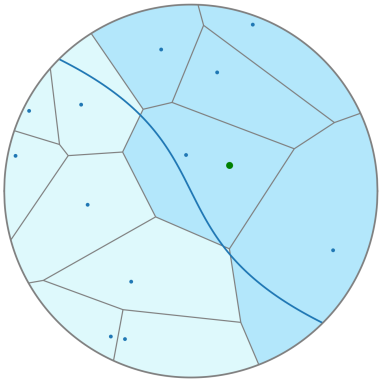
Time	11
Mistake counter	1

I.I.D. sequence



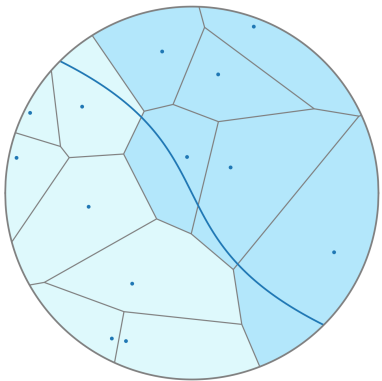
Time	11
Mistake counter	1

I.I.D. sequence



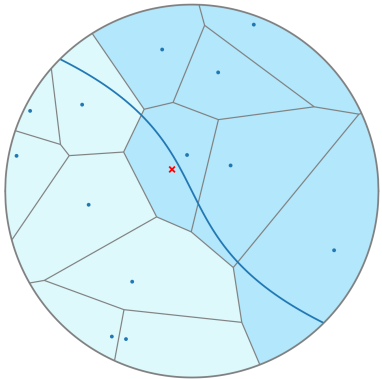
Time	12
Mistake counter	1

I.I.D. sequence



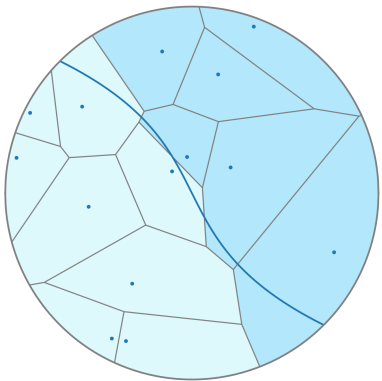
Time	12
Mistake counter	1

I.I.D. sequence



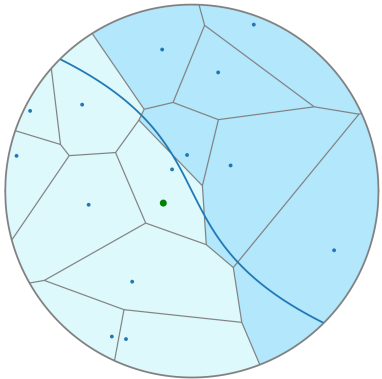
Time	13
Mistake counter	2

I.I.D. sequence



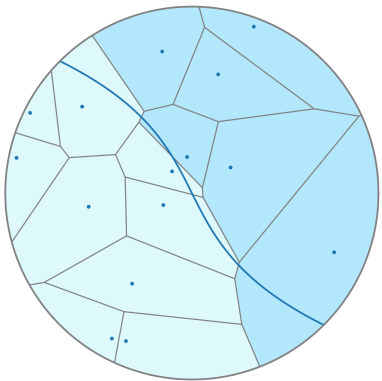
Time	13
Mistake counter	2

I.I.D. sequence



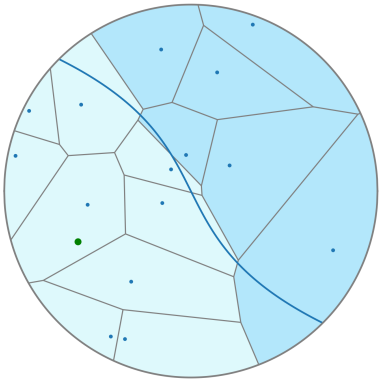
Time	14
Mistake counter	2

I.I.D. sequence



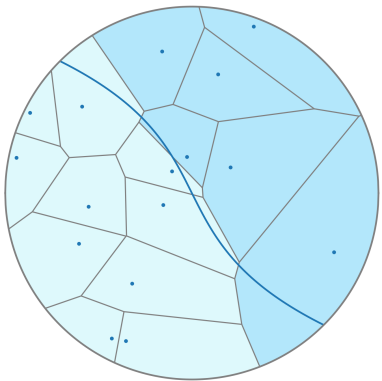
Time	14
Mistake counter	2

I.I.D. sequence



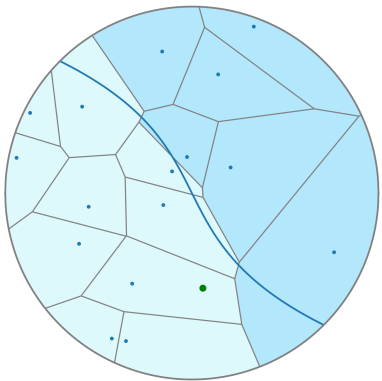
Time	15
Mistake counter	2

I.I.D. sequence



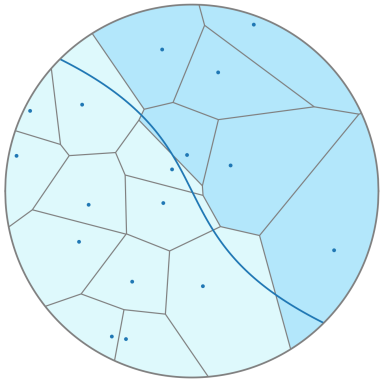
Time	15
Mistake counter	2

I.I.D. sequence



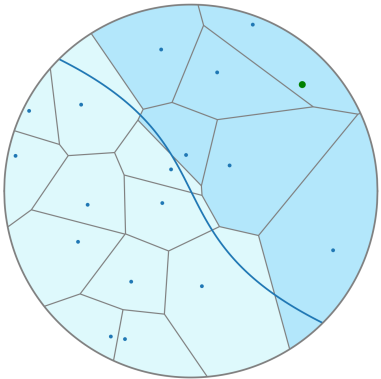
Time	16
Mistake counter	2

I.I.D. sequence



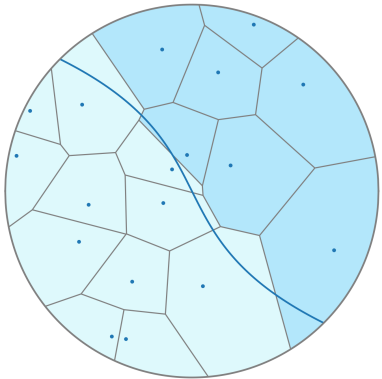
Time	16
Mistake counter	2

I.I.D. sequence



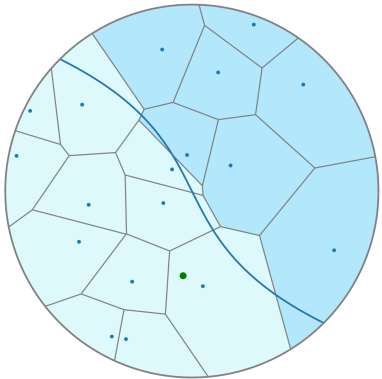
Time	17
Mistake counter	2

I.I.D. sequence



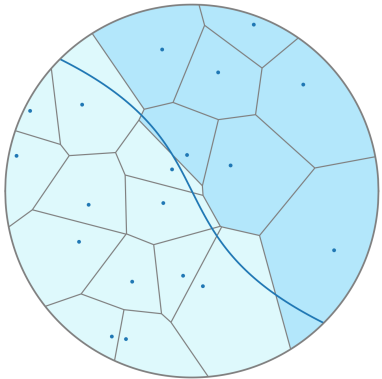
Time	17
Mistake counter	2

I.I.D. sequence



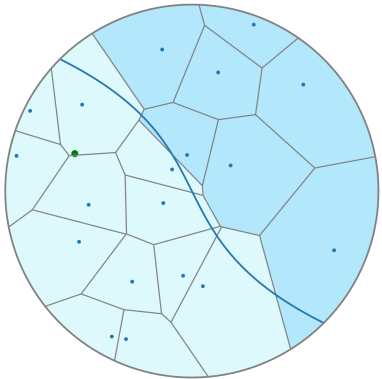
Time	18
Mistake counter	2

I.I.D. sequence



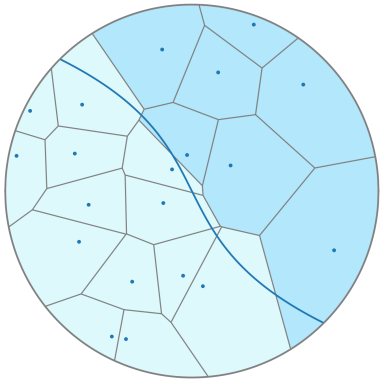
Time	18
Mistake counter	2

I.I.D. sequence



Time	19
Mistake counter	2

I.I.D. sequence

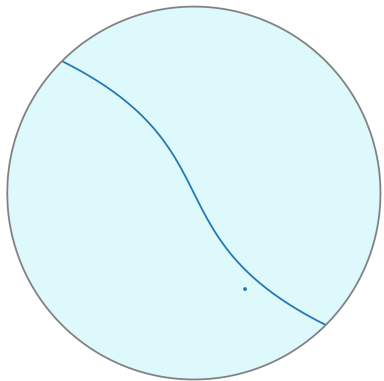


Time	19
Mistake counter	2

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

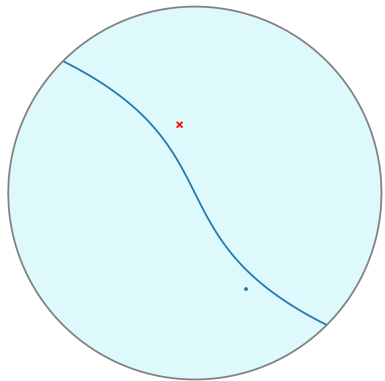
Worst-case sequence

Time	0
Mistake counter	0



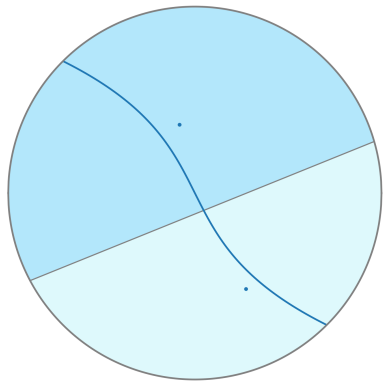
Worst-case sequence

Time	1
Mistake counter	1



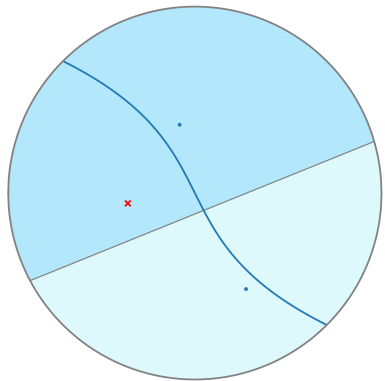
Worst-case sequence

Time	1
Mistake counter	1



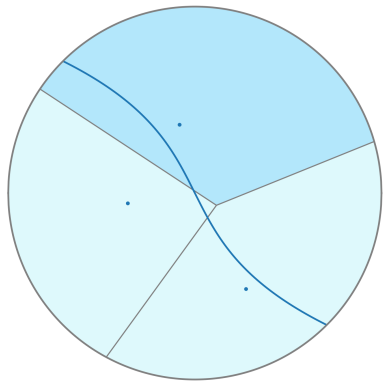
Worst-case sequence

Time	2
Mistake counter	2



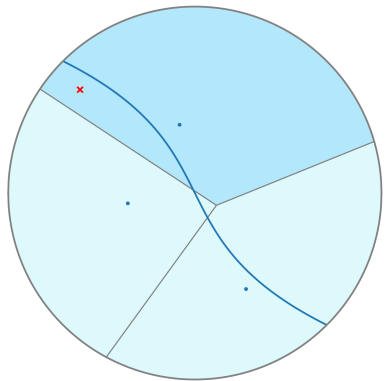
Worst-case sequence

Time	2
Mistake counter	2



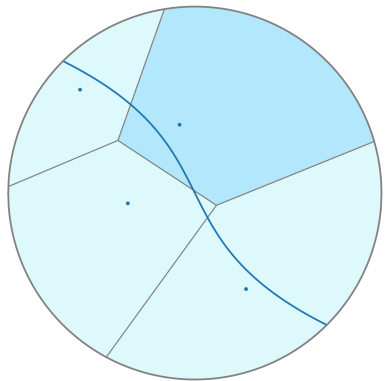
Worst-case sequence

Time	3
Mistake counter	3



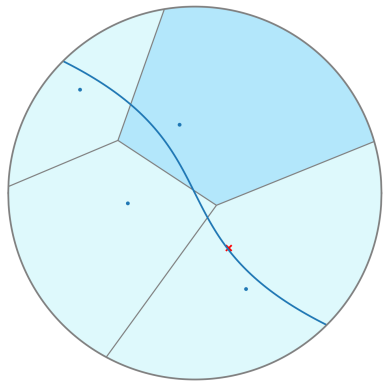
Worst-case sequence

Time	3
Mistake counter	3



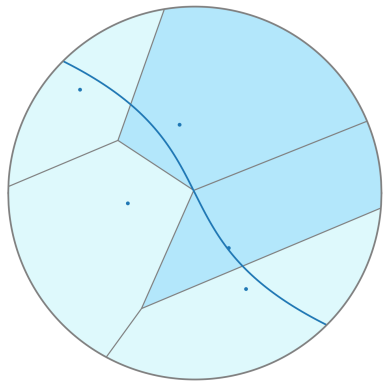
Worst-case sequence

Time	4
Mistake counter	4



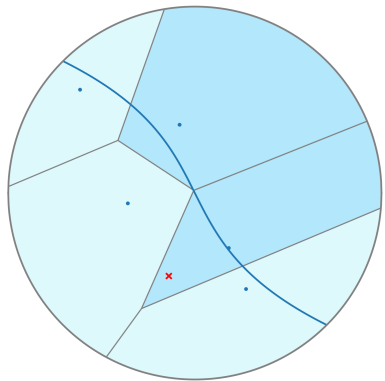
Worst-case sequence

Time	4
Mistake counter	4



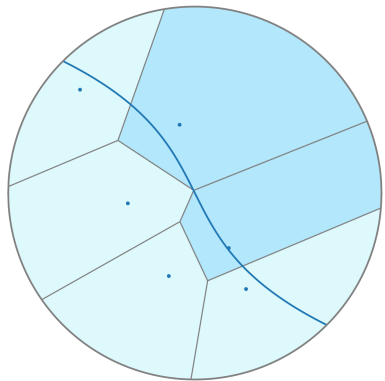
Worst-case sequence

Time	5
Mistake counter	5



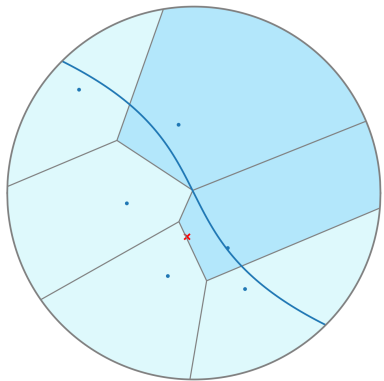
Worst-case sequence

Time	5
Mistake counter	5



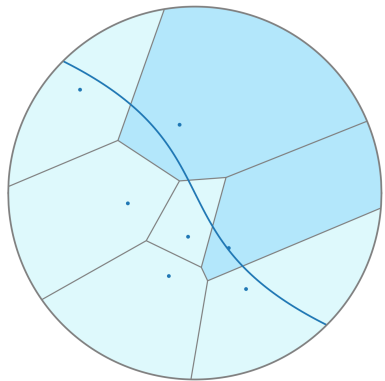
Worst-case sequence

Time	6
Mistake counter	6



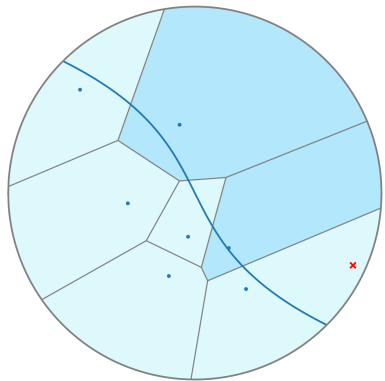
Worst-case sequence

Time	6
Mistake counter	6



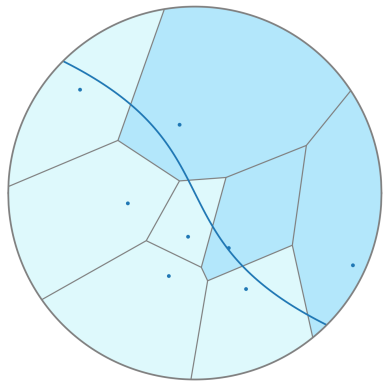
Worst-case sequence

Time	7
Mistake counter	7



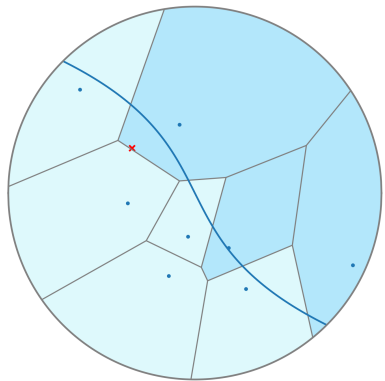
Worst-case sequence

Time	7
Mistake counter	7



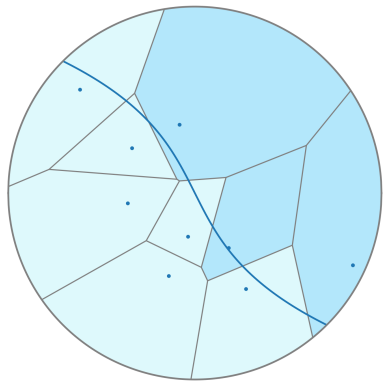
Worst-case sequence

Time	8
Mistake counter	8



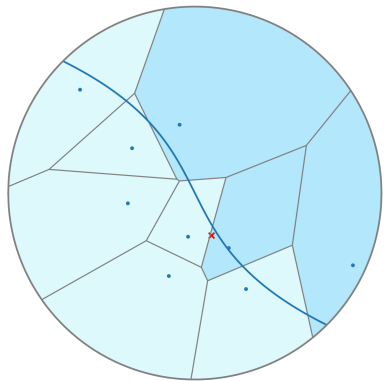
Worst-case sequence

Time	8
Mistake counter	8



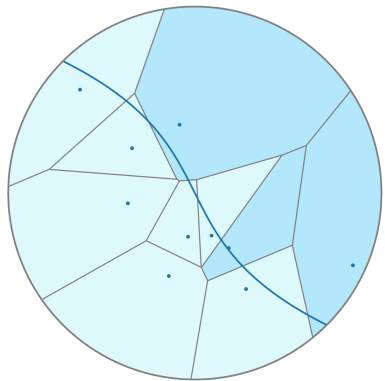
Worst-case sequence

Time	9
Mistake counter	9



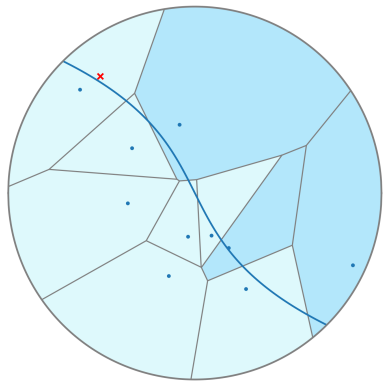
Worst-case sequence

Time	9
Mistake counter	9



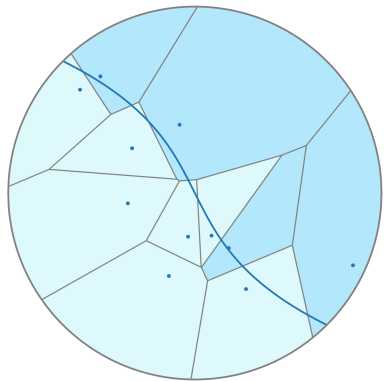
Worst-case sequence

Time	10
Mistake counter	10



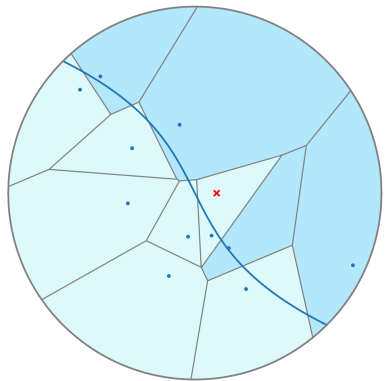
Worst-case sequence

Time	10
Mistake counter	10



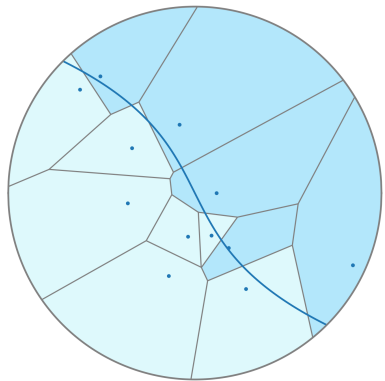
Worst-case sequence

Time	11
Mistake counter	11



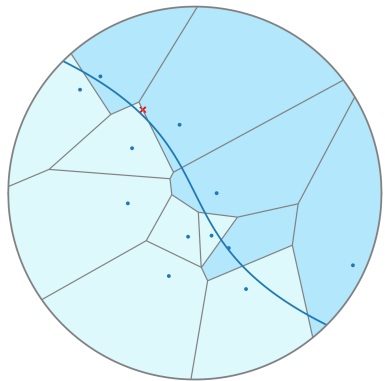
Worst-case sequence

Time	11
Mistake counter	11



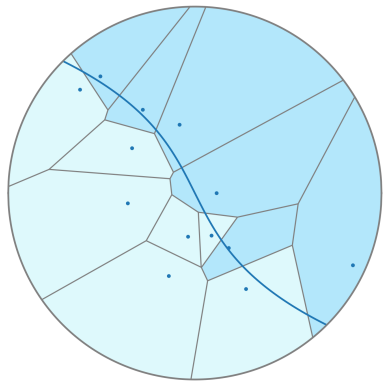
Worst-case sequence

Time	12
Mistake counter	12



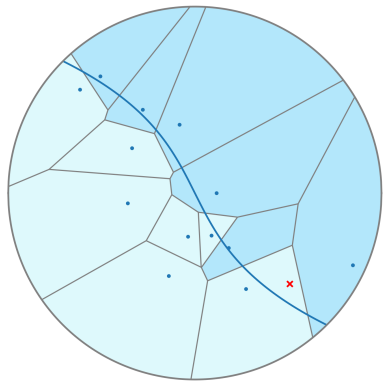
Worst-case sequence

Time	12
Mistake counter	12



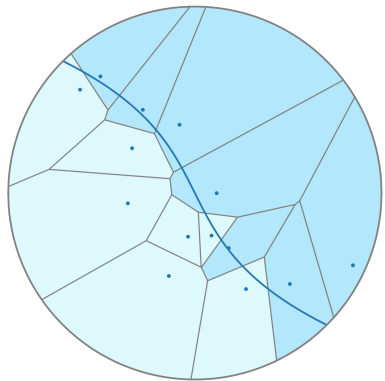
Worst-case sequence

Time	13
Mistake counter	13



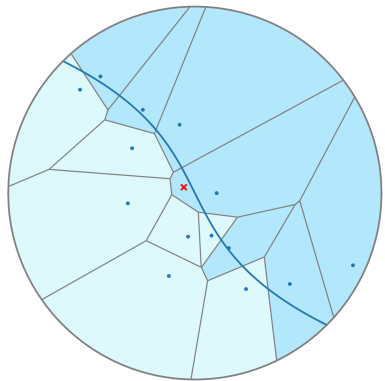
Worst-case sequence

Time	13
Mistake counter	13



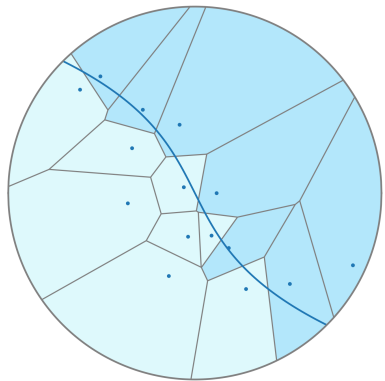
Worst-case sequence

Time	14
Mistake counter	14



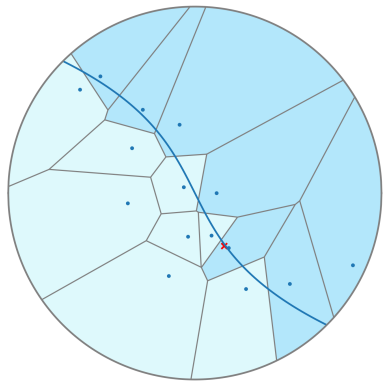
Worst-case sequence

Time	14
Mistake counter	14



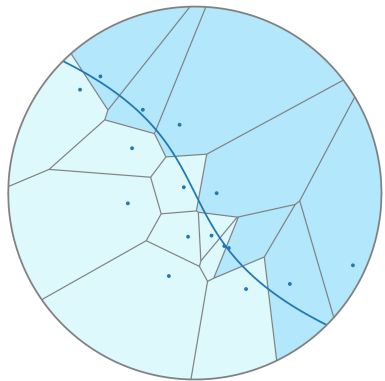
Worst-case sequence

Time	15
Mistake counter	15



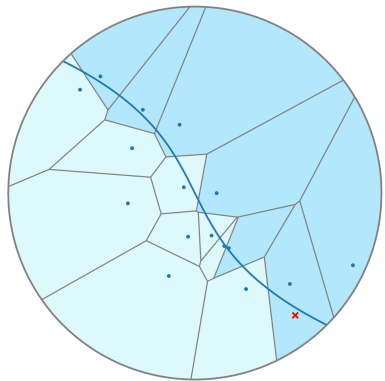
Worst-case sequence

Time	15
Mistake counter	15



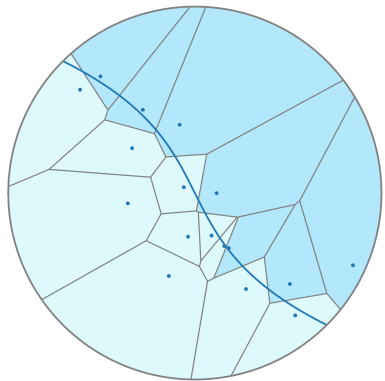
Worst-case sequence

Time	16
Mistake counter	16



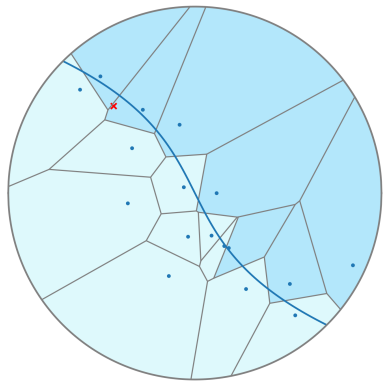
Worst-case sequence

Time	16
Mistake counter	16



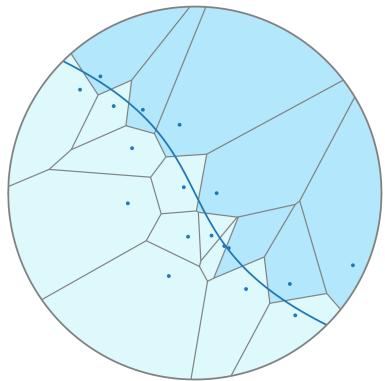
Worst-case sequence

Time	17
Mistake counter	17



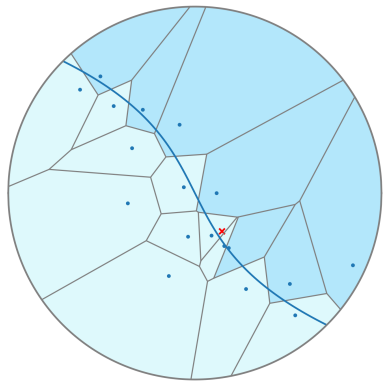
Worst-case sequence

Time	17
Mistake counter	17



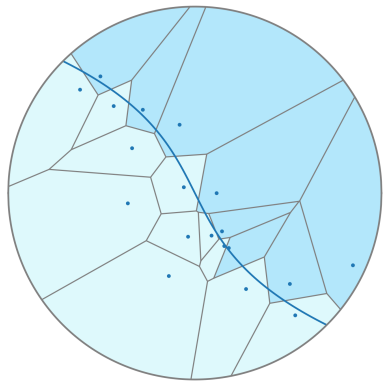
Worst-case sequence

Time	18
Mistake counter	18



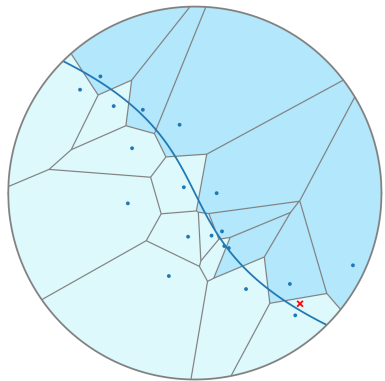
Worst-case sequence

Time	18
Mistake counter	18



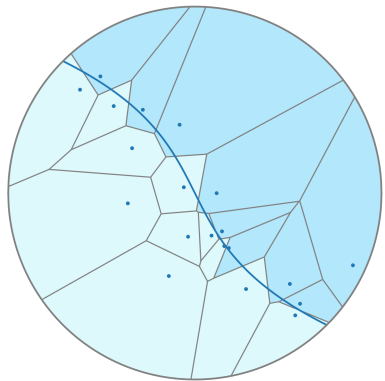
Worst-case sequence

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?

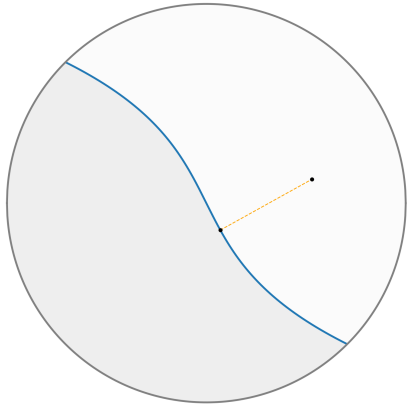
Answer. When different classes have positive separation.

Classification margin

Definition

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

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



Worst-case guarantee for the nearest neighbor rule

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space and $\eta : \mathcal{X} \to \mathcal{Y}$.

Worst-case guarantee for the nearest neighbor rule

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space and $\eta : \mathcal{X} \to \mathcal{Y}$.

Proposition

There exists a sequence $\mathbb{X}$ on which the nearest neighbor rule is not consistent on $(\mathbb{X}, \eta)$

Worst-case guarantee for the nearest neighbor rule

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space and $\eta : \mathcal{X} \to \mathcal{Y}$.

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:

Worst-case guarantee for the nearest neighbor rule

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space and $\eta : \mathcal{X} \to \mathcal{Y}$.

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:

$$\inf_{x \in \mathcal{X}} \operatorname{margin}_{\eta}(x) = 0.$$

Worst-case guarantee for the nearest neighbor rule

($\Longleftarrow$) By contradiction.

Worst-case guarantee for the nearest neighbor rule

($\Longleftarrow$) By contradiction.

- Suppose there is no separation between classes.

Worst-case guarantee for the nearest neighbor rule

($\Longleftarrow$) By contradiction.

- Suppose there is no separation between classes.
- There exists a sequence $\mathbb{X}$ such that:

$$\eta(X_{2k}) \neq \eta(X_{2k+1})$$

Worst-case guarantee for the nearest neighbor rule

($\Longleftarrow$) By contradiction.

- Suppose there is no separation between classes.
- There exists a sequence $\mathbb{X}$ such that:

$$\eta(X_{2k}) \neq \eta(X_{2k+1}) \qquad \text{and} \qquad \tilde{X}_{2k+1} = X_{2k}.$$

Worst-case guarantee for the nearest neighbor rule

($\Longleftarrow$) By contradiction.

- Suppose there is no separation between classes.
- There exists a sequence $\mathbb{X}$ such that:

$$\eta(X_{2k}) \neq \eta(X_{2k+1}) \qquad \text{and} \qquad \tilde{X}_{2k+1} = X_{2k}.$$

- Asymptotic mistake rate is at least $1/2$.

A general class of non-worst case sequences

Excluding the bad cases

Mode of failure.
Provided the learner has not exactly recovered η, a worst-case adversary can always select a point from the mistake region, no matter how small.

Excluding the bad cases

Mode of failure.
Provided the learner has not exactly recovered η, a worst-case adversary can always select a point from the mistake region, no matter how small.

Question.
Can we rule out adversaries that can generate points with arbitrary precision?

Smoothed online setting (Rakhlin et al., 2011)

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

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).
- A test instance X_n is drawn according to μ_n.

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).
- A test instance X_n is drawn according to μ_n.
- The learner makes prediction $\hat{Y}_n$.
- The answer $Y_n = \eta(X_n)$ is revealed.

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).
- A test instance X_n is drawn according to μ_n.
- The learner makes prediction $\hat{Y}_n$.
- The answer $Y_n = \eta(X_n)$ is revealed.

The new addition. Let ν be a reference measure on $\mathcal{X}$.

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).
- A test instance X_n is drawn according to μ_n.
- The learner makes prediction $\hat{Y}_n$.
- The answer $Y_n = \eta(X_n)$ is revealed.

The new addition. Let ν be a reference measure on $\mathcal{X}$.

- We can constrain μ_n with respect to ν (e.g. $\mu_n \ll \nu$).

Smoothed online setting (Rakhlin et al., 2011)

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

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

- A distribution μ_n is chosen (conditioned on $\mathbb{X}_{<n}$).
- A test instance X_n is drawn according to μ_n.
- The learner makes prediction $\hat{Y}_n$.
- The answer $Y_n = \eta(X_n)$ is revealed.

The new addition. Let ν be a reference measure on $\mathcal{X}$.

- We can constrain μ_n with respect to ν (e.g. $\mu_n \ll \nu$).
- Intuitively, it should be hard to generate points from regions with small ν-mass.

Uniform absolute continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ *is* ***uniformly dominated*** *by* ν *if*

Uniform absolute continuity

Definition

*A stochastic process $\mathbb{X} = (X_n)_n$ is **uniformly dominated** by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:*

$$\nu(A) < \delta$$

Uniform absolute continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ is uniformly dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:

$$\nu(A) < \delta \qquad \Longrightarrow$$

Uniform absolute continuity

Definition

*A stochastic process $\mathbb{X} = (X_n)_n$ is **uniformly dominated** by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:*

$$\nu(A) < \delta \qquad \Longrightarrow \qquad \sup_{n \in \mathbb{N}} \underbrace{\Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right)}_{\mu_n(A)} < \varepsilon.$$

Uniform absolute continuity

Definition

*A stochastic process $\mathbb{X} = (X_n)_n$ is **uniformly dominated** by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:*

$$\nu(A) < \delta \qquad \Longrightarrow \qquad \sup_{n \in \mathbb{N}} \underbrace{\Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right)}_{\mu_n(A)} < \varepsilon.$$

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

Uniform absolute continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ is uniformly dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:

$$\nu(A) < \delta \qquad \Longrightarrow \qquad \sup_{n \in \mathbb{N}} \underbrace{\Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right)}_{\mu_n(A)} < \varepsilon.$$

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

Intuition

If $A \subset \mathcal{X}$ has small ν-mass, then the generating process for $\mathbb{X}$ cannot select instances from A with large probability at any point in time.

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$.

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$. Let $(A_n)_n$ be a sequence of subsets such that $\nu(A_n) < \delta$.

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$. Let $(A_n)_n$ be a sequence of subsets such that $\nu(A_n) < \delta$. Then:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A_n\} \leq \varepsilon(\delta) \qquad \text{a.s.}$$

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$. Let $(A_n)_n$ be a sequence of subsets such that $\nu(A_n) < \delta$. Then:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A_n\} \leq \varepsilon(\delta) \qquad \text{a.s.}$$

Intuition (Martingale law of large numbers)

Suppose $\{X_n \in A_n\}$ is a 'bad' or 'atypical' event (e.g. A_n is the mistake region at time n).

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$. Let $(A_n)_n$ be a sequence of subsets such that $\nu(A_n) < \delta$. Then:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A_n\} \leq \varepsilon(\delta) \qquad \text{a.s.}$$

Intuition (Martingale law of large numbers)

Suppose $\{X_n \in A_n\}$ is a 'bad' or 'atypical' event (e.g. A_n is the mistake region at time n).

- *Uniform absolute continuity: it is hard to generate these bad points.*

Arguments via uniform absolute continuity

Proposition

Let $\mathbb{X}$ be uniformly absolutely continuous with respect to ν at rate $\varepsilon(\delta)$. Let $(A_n)_n$ be a sequence of subsets such that $\nu(A_n) < \delta$. Then:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1\{X_n \in A_n\} \leq \varepsilon(\delta) \qquad \text{a.s.}$$

Intuition (Martingale law of large numbers)

Suppose $\{X_n \in A_n\}$ is a 'bad' or 'atypical' event (e.g. A_n is the mistake region at time n).

- *Uniform absolute continuity: it is hard to generate these bad points.*
- *This proposition: on average, we will not see these points very often.*

Consistency for functions with negligible boundaries

Observation.
The nearest neighbor rule works well away from the decision boundary.

Observation.
The nearest neighbor rule works well away from the decision boundary.

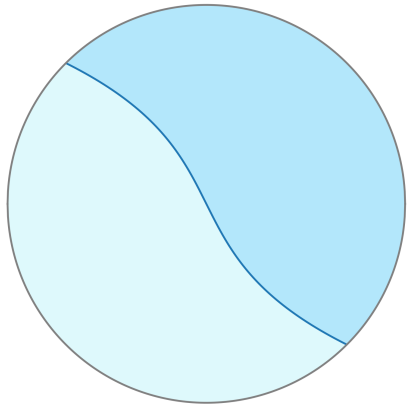
This section.
The nearest neighbor rule has no problem learning functions when boundary points have ν-measure zero and when $\mathbb{X}$ is uniformly dominated by ν.

Functions with negligible boundaries

Definition

Let $\eta : \mathcal{X} \to \mathcal{Y}$. The boundary of η is the set:

$$\partial_\eta \mathcal{X} = \{x \in \mathcal{X} : \text{margin}_\eta(x) = 0\}.$$



Functions with negligible boundaries

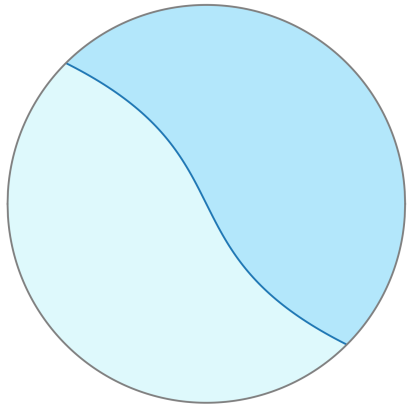
Definition

Let $\eta : \mathcal{X} \to \mathcal{Y}$. The boundary of η is the set:

$$\partial_\eta \mathcal{X} = \{x \in \mathcal{X} : \text{margin}_\eta(x) = 0\}.$$

Denote functions with negligible boundaries:

$$\mathcal{F}_0 = \{\eta \text{ measurable} : \nu(\partial_\eta \mathcal{X}) = 0\}.$$



Functions with negligible boundaries

Definition

Let $\eta : \mathcal{X} \to \mathcal{Y}$. The boundary of η is the set:

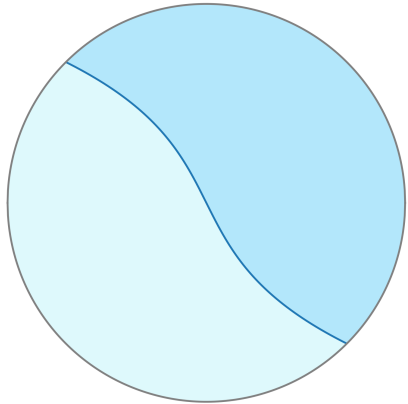
$$\partial_\eta \mathcal{X} = \{x \in \mathcal{X} : \text{margin}_\eta(x) = 0\}.$$

Denote functions with negligible boundaries:

$$\mathcal{F}_0 = \{\eta \text{ measurable} : \nu(\partial_\eta \mathcal{X}) = 0\}.$$

Example

Let $\eta : \mathbb{R}^d \to \{0, 1\}$ have a smooth decision boundary, which is a $(d-1)$-dimensional object in $\mathbb{R}^d$, and thus has zero Lebesgue measure.



Consistency for $\mathcal{F}_0$

Theorem

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

Consistency for $\mathcal{F}_0$

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 uniformly dominated by ν and η has negligible boundary.

Consistency for $\mathcal{F}_0$

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 uniformly dominated by ν and η has negligible boundary. Then:

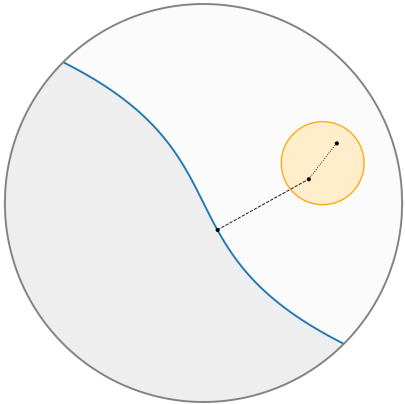
$$\underbrace{\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\eta(X_n) \neq \eta(\tilde{X}_n)\} = 0}_{\text{the nearest neighbor rule is online consistent for } (\mathbb{X}, \eta).} \qquad \text{a.s.}$$

Mutually-labeling set

Definition

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

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



Mutually-labeling set

Definition

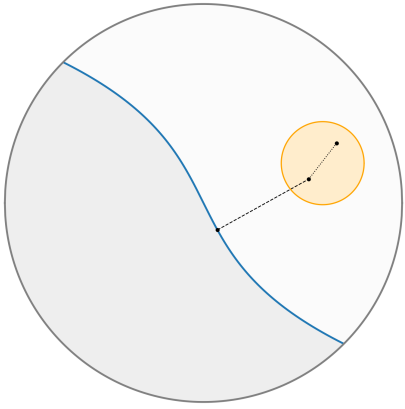
*A set $U \subset \mathcal{X}$ is **mutually-labeling** for η when:*

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

Proposition

When x is not a boundary point, the ball $B(x, r)$ is mutually labeling when:

$$r < \operatorname{margin}_\eta(x)/3.$$



Mutually-labeling property

Definition

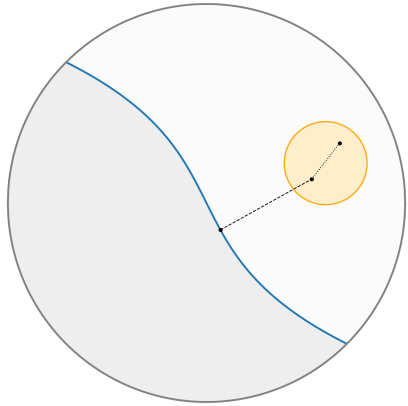
*A set $U \subset \mathcal{X}$ is **mutually-labeling** for η when:*

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

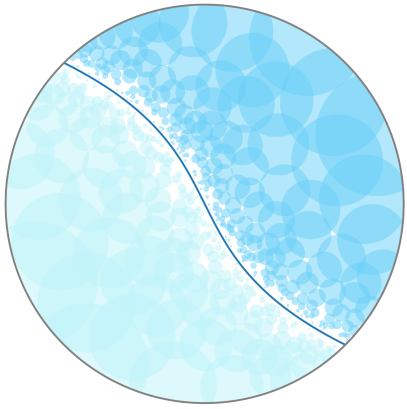
Proposition

Let U be a mutually-labeling set for η. The nearest neighbor rule makes at most one mistake on U,

$$\sum_{n=1}^{\infty} \mathbb{1}\{\eta(X_n) \neq \eta(\tilde{X}_n) \wedge X_n \in U\} \leq 1.$$

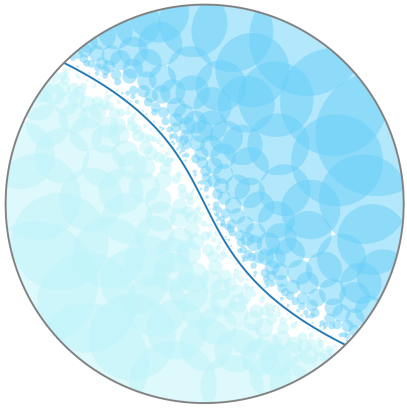


Proof by a δ-approximate mutually-labeling cover



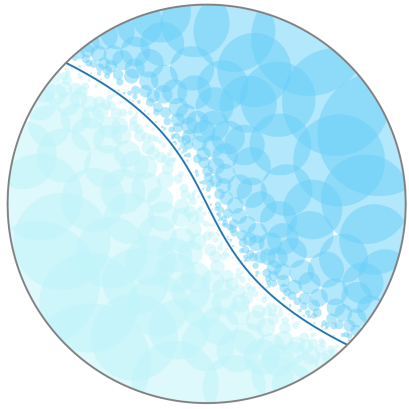
Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.



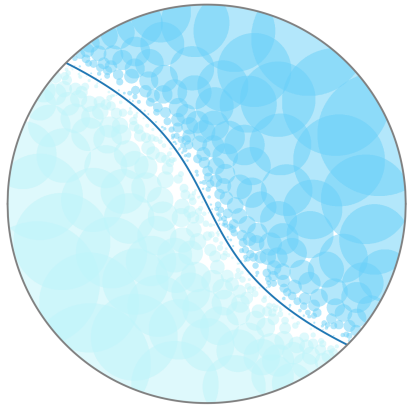
Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.



Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.
3. As $\mathbb{X}$ is uniformly dominated, the rate at which this occurs is at most $\varepsilon(\delta)$.

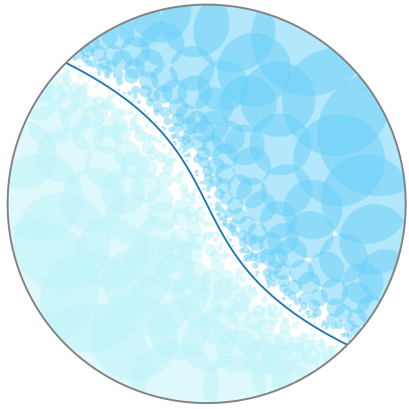


Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.
3. As $\mathbb{X}$ is uniformly dominated, the rate at which this occurs is at most $\varepsilon(\delta)$.

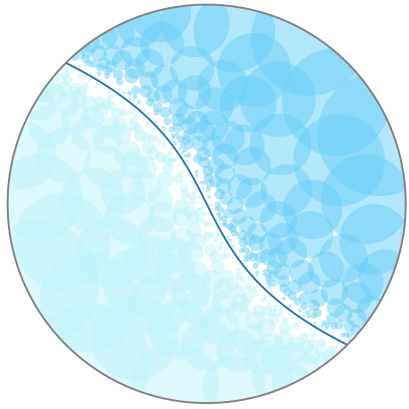
Construction

Consider all mutually-labeling open balls.



Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.
3. As $\mathbb{X}$ is uniformly dominated, the rate at which this occurs is at most $\varepsilon(\delta)$.



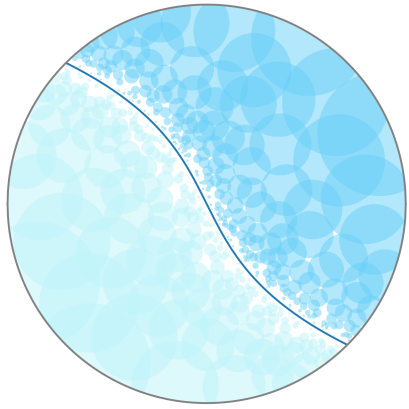
Construction

Consider all mutually-labeling open balls.

- *We only miss the null set $\partial_\eta \mathcal{X}$ (η is in $\mathcal{F}_0$).*

Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.
3. As $\mathbb{X}$ is uniformly dominated, the rate at which this occurs is at most $\varepsilon(\delta)$.



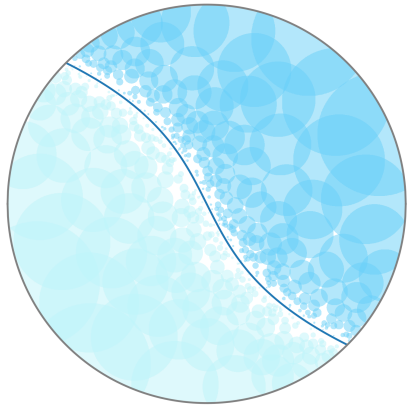
Construction

Consider all mutually-labeling open balls.

- *We only miss the null set $\partial_\eta \mathcal{X}$ (η is in $\mathcal{F}_0$).*
- *Take a countable subcover (ρ is separable).*

Proof by a δ-approximate mutually-labeling cover

1. Find a finite collection of mutually-labeling sets covering all but a region of ν-mass δ.
2. Note that eventually all mistakes must come from the uncovered region.
3. As $\mathbb{X}$ is uniformly dominated, the rate at which this occurs is at most $\varepsilon(\delta)$.



Construction

Consider all mutually-labeling open balls.

- *We only miss the null set $\partial_\eta \mathcal{X}$ (η is in $\mathcal{F}_0$).*
- *Take a countable subcover (ρ is separable).*
- *Take a finite δ-approx. cover (ν is finite).*

Consistency on upper doubling spaces

Observation.
In Euclidean space, simple functions are dense in L^1.

piecewise constant

Observation.
In Euclidean space, simple functions are dense in L^1.

piecewise constant

This section.
It turns out that $\mathcal{F}_0$ is dense in L^1 for nice metric measure spaces.

Observation.
In Euclidean space, simple functions are dense in L^1.

piecewise constant

This section.
It turns out that $\mathcal{F}_0$ is dense in L^1 for nice metric measure spaces.

The nearest neighbor rule behaves similarly for similar functions $\eta \approx \eta'$

Observation.
In Euclidean space, simple functions are dense in L^1.

piecewise constant

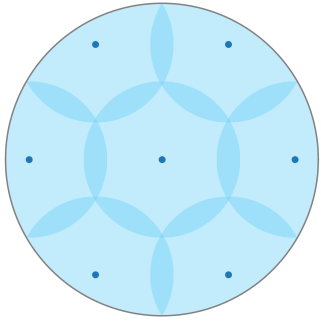
This section.
It turns out that $\mathcal{F}_0$ is dense in L^1 for nice metric measure spaces.

The nearest neighbor rule behaves similarly for similar functions $\eta \approx \eta'$ when $\mathbb{X}$ is uniformly dominated in a geometrically nice space.

Upper doubling spaces

Definition (Hytönen (2010))

A metric measure space $(\mathcal{X}, \rho, \nu)$ is upper doubling when:

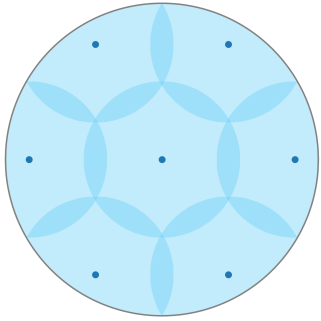


Upper doubling spaces

Definition (Hytönen (2010))

*A metric measure space $(\mathcal{X}, \rho, \nu)$ is **upper doubling** when:*

1. *The metric ρ is doubling with doubling dimension d.*



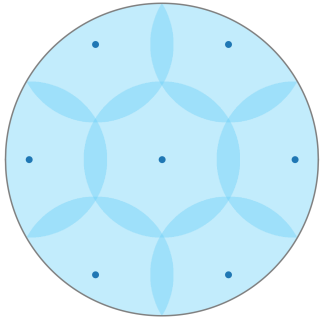
Upper doubling spaces

Definition (Hytönen (2010))

A metric measure space $(\mathcal{X}, \rho, \nu)$ is upper doubling when:

1. *The metric ρ is doubling with doubling dimension d. That is, the covering number is bounded:*

$$\mathcal{N}_{r/2}\big(B(x,r)\big) \leq 2^d.$$



Upper doubling spaces

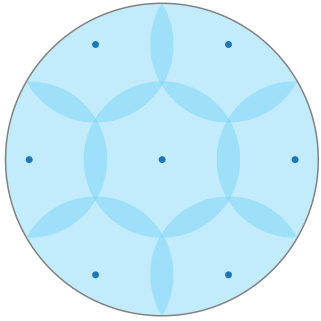
Definition (Hytönen (2010))

A metric measure space $(\mathcal{X}, \rho, \nu)$ is upper doubling when:

1. *The metric ρ is doubling with doubling dimension d. That is, the covering number is bounded:*

$$\mathcal{N}_{r/2}\big(B(x,r)\big) \leq 2^d.$$

2. *The measure ν is upper doubling.*



Upper doubling spaces

Definition (Hytönen (2010))

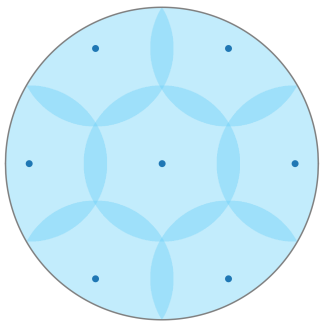
A metric measure space $(\mathcal{X}, \rho, \nu)$ is upper doubling when:

1. *The metric ρ is doubling with doubling dimension d. That is, the covering number is bounded:*

$$\mathcal{N}_{r/2}\big(B(x,r)\big) \leq 2^d.$$

2. *The measure ν is upper doubling. That is, there exists a constant $c > 0$ where:*

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



$\mathcal{F}_0$ is dense in $L^1(\mathcal{X}; \nu)$

Let $\mathcal{X}$ be equipped with a doubling metric and a finite measure.

Question. Can we approximate η arbitrarily well by $\eta_0 \in \mathcal{F}_0$? Yes.

Lebesgue differentiation theorem. For almost every $x \in \mathcal{X}$, we can find arbitrarily small balls $B(x, r)$ such that η is nearly constant on $B(x, r)$.

Vitali covering theorem. There is a countable disjoint subcollection of such balls $B(x_1, r_1), B(x_2, r_2), \ldots$ that cover almost all of $\mathcal{X}$.

Construction. Define $\eta_0 = \sum_{i=1}^{\infty} \eta(x_i) \cdot \mathbb{1}_{B(x_i, r_i)}$, where $B(x_i, r_i) \cap \partial_{\eta_0} \mathcal{X} = \varnothing$.

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 strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

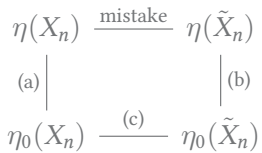
2. Argue that the nearest neighbor rule behaves similarly on η and η_0.

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

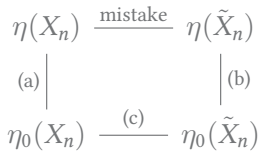
Given a mistake on (X_n, η), at least one must occur:

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

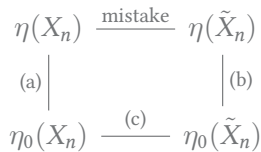
(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

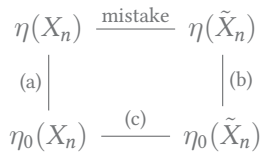
(b) *$\tilde{X}_n$ is from the disagreement region $\{\eta \neq \eta_0\}$.*

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

(b) *$\tilde{X}_n$ is from the disagreement region $\{\eta \neq \eta_0\}$.*

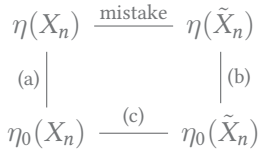
(c) *The nearest neighbor rule errs on (X_n, η_0).*

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

(b) *$\tilde{X}_n$ is from the disagreement region $\{\eta \neq \eta_0\}$.*

(c) *The nearest neighbor rule errs on (X_n, η_0).*

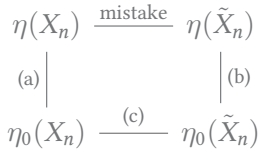
As $\mathbb{X}$ is uniformly dominated, this occurs at rate at most $\varepsilon(\delta)$.

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

(b) *$\tilde{X}_n$ is from the disagreement region $\{\eta \neq \eta_0\}$.*

(c) *The nearest neighbor rule errs on (X_n, η_0).*

As $\mathbb{X}$ is uniformly dominated, this occurs at rate at most $\varepsilon(\delta)$.

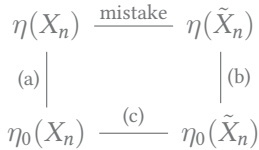
By consistency on $\mathcal{F}_0$, the rate for this term vanishes.

Proof strategy

1. Approximate η by $\eta_0 \in \mathcal{F}_0$ such that:

$$\nu\big(\{\eta \neq \eta_0\}\big) < \delta.$$

2. Argue that the nearest neighbor rule behaves similarly on η and η_0.



Argument

Given a mistake on (X_n, η), at least one must occur:

(a) *X_n is from the disagreement region $\{\eta \neq \eta_0\}$.*

(b) *$\tilde{X}_n$ is from the disagreement region $\{\eta \neq \eta_0\}$.*

(c) *The nearest neighbor rule errs on (X_n, η_0).*

As $\mathbb{X}$ is uniformly dominated, this occurs at rate at most $\varepsilon(\delta)$.

$\tilde{\mathbb{X}}$ is ergodically dominated, so we can control this rate as well.

By consistency on $\mathcal{F}_0$, the rate for this term vanishes.

Ergodic continuity of nearest neighbor processes

Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?

Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?

generating process: X_0

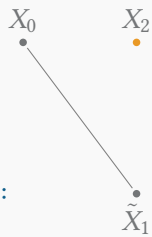
nearest neighbor process:

Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?

generating process:

nearest neighbor process:

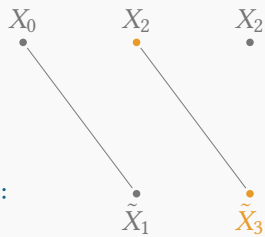


Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?

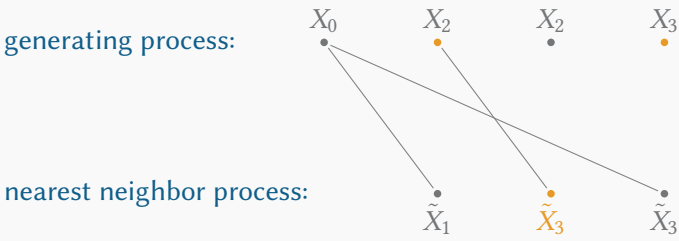
generating process:

nearest neighbor process:



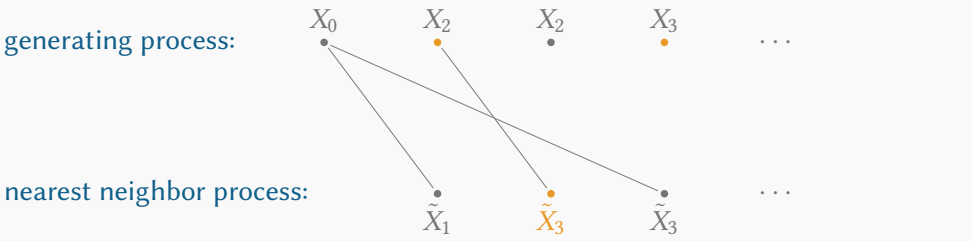
Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?



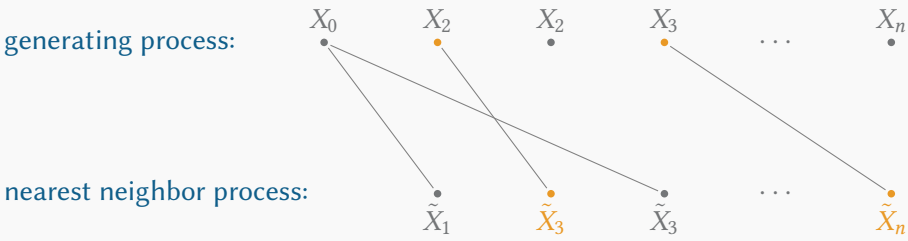
Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?



Influence filtered through the nearest neighbor process

Question. For any $A \subset \mathcal{X}$, what is the long-term influence of points in $\mathbb{X} \cap A$ on $\tilde{\mathbb{X}}$?



Two opposing dynamics

Goal. Upper bound the asymptotic rate that $\tilde{\mathbb{X}}$ hits A:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\}.$$

Two opposing dynamics

Goal. Upper bound the asymptotic rate that $\tilde{\mathbb{X}}$ hits A:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\}.$$

1. Positive dynamics. More instances X_n can accumulate in A over time.

Two opposing dynamics

Goal. Upper bound the asymptotic rate that $\tilde{\mathbb{X}}$ hits A:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\}.$$

1. Positive dynamics. More instances X_n can accumulate in A over time.
2. Negative dynamics. Voronoi cells of past instances shrink when hit.

Two opposing dynamics

Goal. Upper bound the asymptotic rate that $\tilde{\mathbb{X}}$ hits A:

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\}.$$

1. Positive dynamics. More instances X_n can accumulate in A over time.
2. Negative dynamics. Voronoi cells of past instances shrink when hit.

Upshot. These two dynamics balance out when $\mathbb{X}$ is a uniformly dominated process in an upper doubling space.

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$
- $\mathcal{P}_r(U)$ is the r-packing number of U

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$
- $\mathcal{P}_r(U)$ is the r-packing number of U
- $\mathbb{X}$ is an arbitrary process on $\mathcal{X}$

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$
- $\mathcal{P}_r(U)$ is the r-packing number of U
- $\mathbb{X}$ is an arbitrary process on $\mathcal{X}$

Lemma

The number of times that $\mathbb{X}$ *hits* U

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$
- $\mathcal{P}_r(U)$ is the r-packing number of U
- $\mathbb{X}$ is an arbitrary process on $\mathcal{X}$

Lemma

The number of times that $\mathbb{X}$ hits U while having nearest neighbor distance at least r is:

Analysis of the negative dynamics via a packing argument

Setting.

- $(\mathcal{X}, \rho)$ is any metric space and $U \subset \mathcal{X}$
- $\mathcal{P}_r(U)$ is the r-packing number of U
- $\mathbb{X}$ is an arbitrary process on $\mathcal{X}$

Lemma

The number of times that $\mathbb{X}$ *hits* U *while having nearest neighbor distance at least* r *is:*

$$\sum_{n=1}^{\infty} \mathbb{1}\{ X_n \in U \ \wedge \ \rho(X_n, \tilde{X}_n) \geq r \} \leq \mathcal{P}_r(U).$$

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process.

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process. For any $N, k \in \mathbb{N}$,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\}$$

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process. For any $N, k \in \mathbb{N}$,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\} \leq 2^d \cdot k + \sum_{n=1}^{N} \mathbb{1}\{X_n \in B_0\},$$

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process. For any $N, k \in \mathbb{N}$,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\} \leq 2^d \cdot k + \sum_{n=1}^{N} \mathbb{1}\{X_n \in B_0\},$$

where $B_0 = B(X_0, R \cdot 2^{-k})$ and $R = \operatorname{diam}(\mathcal{X})$.

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process. For any $N, k \in \mathbb{N}$,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\} \leq 2^d \cdot k + \sum_{n=1}^{N} \mathbb{1}\{X_n \in B_0\},$$

where $B_0 = B(X_0, R \cdot 2^{-k})$ and $R = \operatorname{diam}(\mathcal{X})$.

- The first term comes from the packing number of *doubling* spaces.

Warm-up: diminishing influence of a single point

Proposition

Let $\mathcal{X}$ be a bounded doubling space. Let $\mathbb{X}$ be any process. For any $N, k \in \mathbb{N}$,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n = X_0\} \leq 2^d \cdot k + \sum_{n=1}^{N} \mathbb{1}\{X_n \in B_0\},$$

where $B_0 = B(X_0, R \cdot 2^{-k})$ and $R = \operatorname{diam}(\mathcal{X})$.

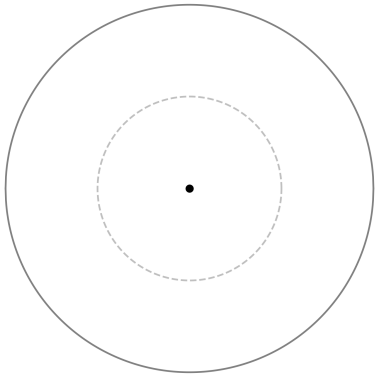
- The first term comes from the packing number of *doubling* spaces.
- The second term can be controlled when the space is *upper* doubling:

$$\nu(B_0) = 2^{-\Omega(k)}.$$

Warm-up: diminishing influence of a single point

X_0 is the nearest neighbor:

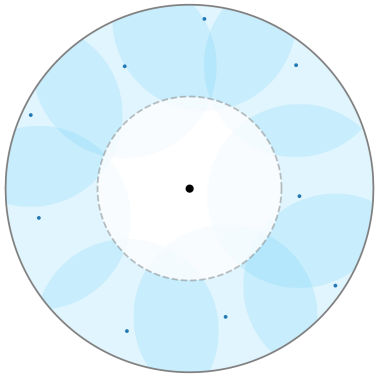
1. at scale $[R/2, R)$ no more than 2^d times.



Warm-up: diminishing influence of a single point

X_0 is the nearest neighbor:

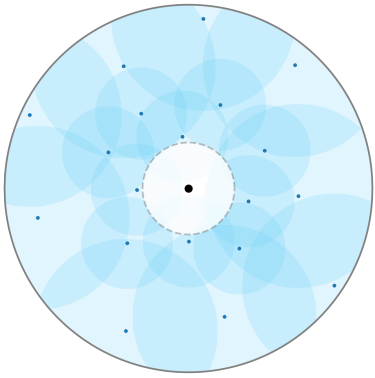
1. at scale $[R/2, R)$ no more than 2^d times.



Warm-up: diminishing influence of a single point

X_0 is the nearest neighbor:

1. at scale $[R/2, R)$ no more than 2^d times.
2. at scale $[R/4, R/2)$ no more than 2^d times.

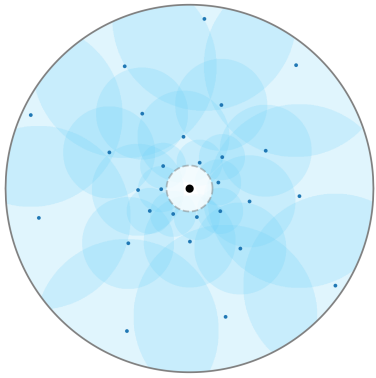


Warm-up: diminishing influence of a single point

X_0 is the nearest neighbor:

1. at scale $[R/2, R)$ no more than 2^d times.
2. at scale $[R/4, R/2)$ no more than 2^d times.
3. at scale $[R/8, R/4)$ no more than 2^d times.

$\vdots$



Warm-up: diminishing influence of a single point

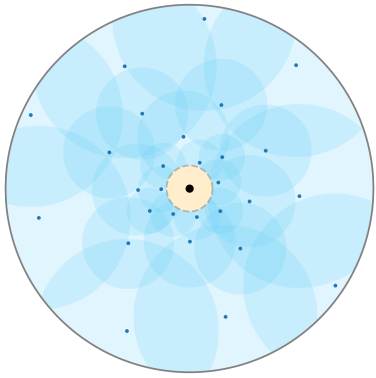
X_0 is the nearest neighbor:

1. at scale $[R/2, R)$ no more than 2^d times.
2. at scale $[R/4, R/2)$ no more than 2^d times.
3. at scale $[R/8, R/4)$ no more than 2^d times.

$\vdots$

If $\mathcal{X}$ is upper doubling and $\mathbb{X}$ is uniformly dominated, then X_0 is the nearest neighbor:

k. at scale $< R \cdot 2^{-k}$ at a rate $2^{-\Omega(k)}$.

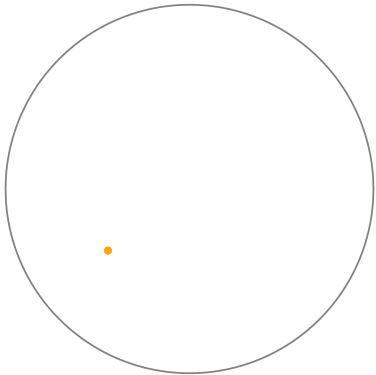


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\cdots$ X_n

- At time $n = 1$, we observe $X_1 \in A$.

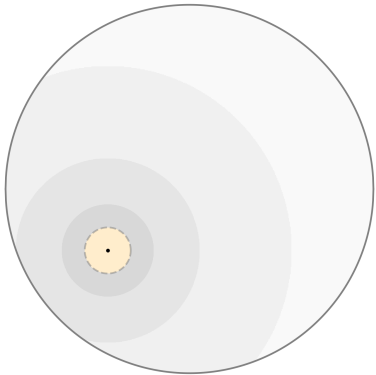


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

- At time $n = 1$, we observe $X_1 \in A$.
- We can pay upfront for part of $\{\tilde{X}_n = X_1\}$.



Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\cdots$ X_n

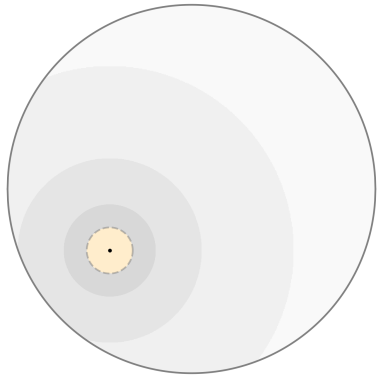
- At time $n = 1$, we observe $X_1 \in A$.
- We can pay upfront for part of $\{\tilde{X}_n = X_1\}$.

Packing Argument.

1. For each ring of outer radius r, the event:

$$\{X_n, \tilde{X}_n \in B(X_1, r) \wedge \rho(X_n, \tilde{X}_n) > r/2\},$$

happens no more than $O(2^d)$ times.



Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

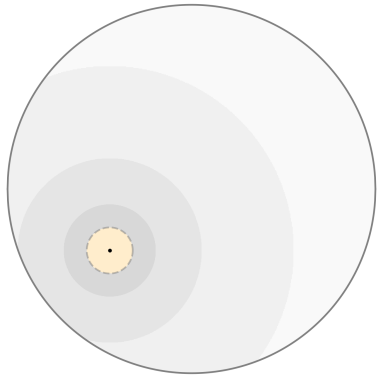
- At time $n = 1$, we observe $X_1 \in A$.
- We can pay upfront for part of $\{\tilde{X}_n = X_1\}$.

Packing Argument.

1. For each ring of outer radius r, the event:

$$\{X_n, \tilde{X}_n \in B(X_1, r) \wedge \rho(X_n, \tilde{X}_n) > r/2\},$$

   happens no more than $O(2^d)$ times.
2. Remainder contained in $\{X_n \in B(X_1, r/2)\}$.

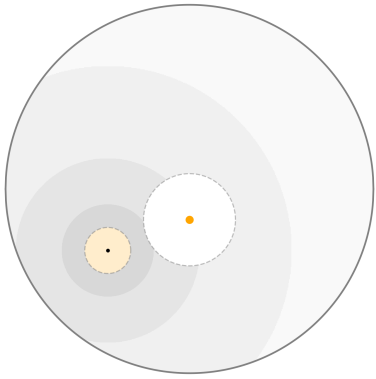


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\cdots$ X_n

- At $n = 3$, we observe $X_3 \in A$.

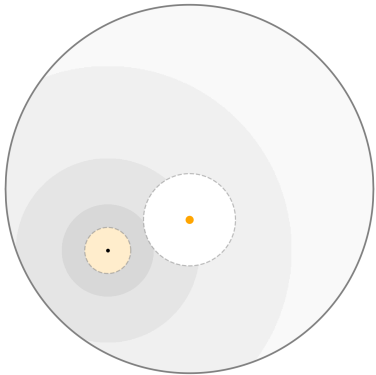


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

- At $n = 3$, we observe $X_3 \in A$.
- We need to account for the event $\{\tilde{X}_n = X_3\}$.

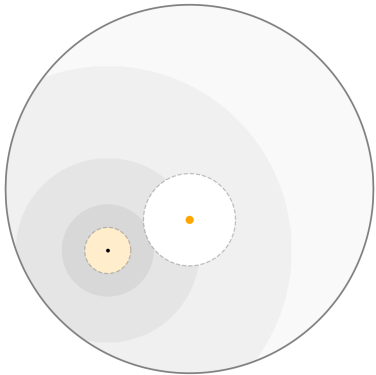


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

- At $n = 3$, we observe $X_3 \in A$.
- We need to account for the event $\{\tilde{X}_n = X_3\}$.
- Part of this event is already accounted for.

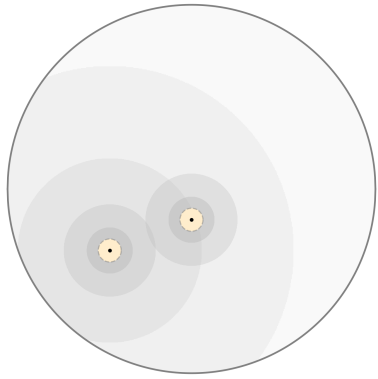


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

- At $n = 3$, we observe $X_3 \in A$.
- We need to account for the event $\{\tilde{X}_n = X_3\}$.
- Part of this event is already accounted for.
- Pay for more scales for both X_3 and parent(X_3).

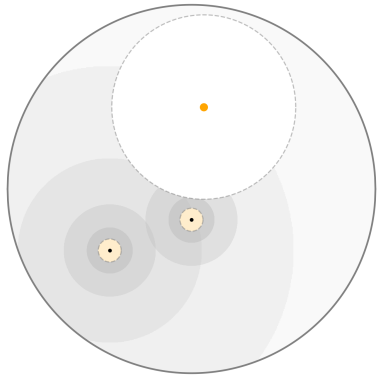


Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\dots$ X_n

- At $n = 3$, we observe $X_3 \in A$.
- We need to account for the event $\{\tilde{X}_n = X_3\}$.
- Part of this event is already accounted for.
- Pay for more scales for both X_3 and $\text{parent}(X_3)$.
- Induct on each X_n that comes from A,



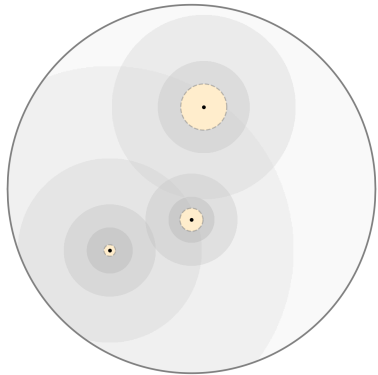
Bounding the long-term influence of a subsequence

Generating process:

X_0 X_1 X_2 X_3 $\cdots$ X_n

- At $n = 3$, we observe $X_3 \in A$.
- We need to account for the event $\{\tilde{X}_n = X_3\}$.
- Part of this event is already accounted for.
- Pay for more scales for both X_3 and $\text{parent}(X_3)$.
- Induct on each X_n that comes from A, ensuring:

$$\nu\big(\text{remainder region}\big) < \delta.$$



Bounding the long-term influence of a subsequence

By this argument, at every point in time N,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\} \leq O\left(\left|\mathbb{X}_{<N} \cap A\right|\right) + \sum_{n=1}^{N} \mathbb{1}\{X_n \in R_n\},$$

Bounding the long-term influence of a subsequence

By this argument, at every point in time N,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\} \leq O\left(\left|\mathbb{X}_{<N} \cap A\right|\right) + \sum_{n=1}^{N} \mathbb{1}\{X_n \in R_n\},$$

where R_n is the sequence of remainder regions with $\nu(R_n) < \delta$.

Bounding the long-term influence of a subsequence

By this argument, at every point in time N,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\} \leq O\left(\left|\mathbb{X}_{<N} \cap A\right|\right) + \sum_{n=1}^{N} \mathbb{1}\{X_n \in R_n\},$$

where R_n is the sequence of remainder regions with $\nu(R_n) < \delta$.

- The first term grows at rate depending on $O\left(\varepsilon(\delta) \log \frac{1}{\delta}\right)$.

Bounding the long-term influence of a subsequence

By this argument, at every point in time N,

$$\sum_{n=1}^{N} \mathbb{1}\{\tilde{X}_n \in A\} \leq O\left(\left|\mathbb{X}_{<N} \cap A\right|\right) + \sum_{n=1}^{N} \mathbb{1}\{X_n \in R_n\},$$

where R_n is the sequence of remainder regions with $\nu(R_n) < \delta$.

- The first term grows at rate depending on $O\left(\varepsilon(\delta) \log \frac{1}{\delta}\right)$.
- The second term grows at a rate bounded by $\varepsilon(\delta)$.

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. Suppose $\mathbb{X}$ is uniformly dominated by ν at a rate of $\varepsilon(\delta)$.

Ergodic continuity of the nearest neighbor process

Theorem

Let $(\mathcal{X}, \rho, \nu)$ be bounded and upper doubling. Suppose $\mathbb{X}$ is uniformly dominated by ν at a rate of $\varepsilon(\delta)$. Then, $\tilde{\mathbb{X}}$ is ergodically dominated by ν at a slower rate of $O\big(\varepsilon(\delta)\log\frac{1}{\delta}\big)$.

Ergodic continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ *is* ***ergodically dominated*** *by* ν

Ergodic continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ is ergodically dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:

$$\nu(A) < \delta$$

Ergodic continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ is ergodically dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:

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

Ergodic continuity

Definition

A stochastic process $\mathbb{X} = (X_n)_n$ is ergodically dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all measurable $A \subset \mathcal{X}$:

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

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

Intuition

If $A \subset \mathcal{X}$ has small ν-mass, then the generating process for $\mathbb{X}$ cannot have selected instances from A very often in retrospect.

Uniform absolute continuity vs. ergodic continuity

Setting. Let $A \subset \mathcal{X}$ have measure $\nu(A) < \delta$.

Uniform absolute continuity vs. ergodic continuity

Setting. Let $A \subset \mathcal{X}$ have measure $\nu(A) < \delta$.

Uniform absolute continuity

$$\underbrace{\sup_{n \in \mathbb{N}} \Pr\left(X_n \in A\right) < \varepsilon}_{\text{time-uniform guarantee}}$$

Uniform absolute continuity vs. ergodic continuity

Setting. Let $A \subset \mathcal{X}$ have measure $\nu(A) < \delta$.

Uniform absolute continuity

$$\underbrace{\sup_{n \in \mathbb{N}} \Pr\left(X_n \in A\right) < \varepsilon}_{\text{time-uniform guarantee}} \implies$$

Uniform absolute continuity vs. ergodic continuity

Setting. Let $A \subset \mathcal{X}$ have measure $\nu(A) < \delta$.

Uniform absolute continuity

$$\underbrace{\sup_{n \in \mathbb{N}} \Pr\left(X_n \in A\right) < \varepsilon}_{\text{time-uniform guarantee}}$$

$\Longrightarrow$

Ergodic continuity

$$\underbrace{\limsup_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{X_n \in A\} < \varepsilon}_{\text{time-averaged guarantee}}$$

Upshot

Suppose that $(\mathcal{X}, \rho, \nu)$ is a reasonable space, and test instances are generated by an adversary with time-uniform bounded precision.

Then, the nearest neighbor rule will eventually make no mistakes on average.

Looking forward

Immediate follow-up questions

1. Online consistency of the k_n-nearest neighbor rule for noisy labels
2. Nearest neighbor methods for adversarial labels
3. Learning with bounded or rate-limited memory
4. Extension to general distance spaces
5. Comparison of uniform absolute continuity to weaker conditions in literature

Some other ongoing work

- *Learning in repeated games with a shared signal*, with Yian Ma and Yixin Wang.
- *Optimization on the Pareto set*, with Yian Ma and Abhishek Roy.

Thank you!

Additional slides

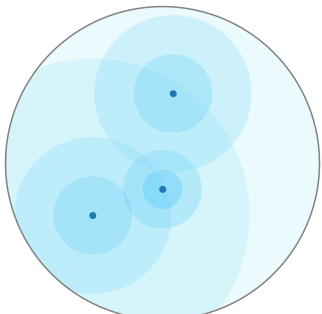
Cover tree data structure

Let $\mathcal{D}$ be a finite dataset in $(\mathcal{X}, \rho)$.

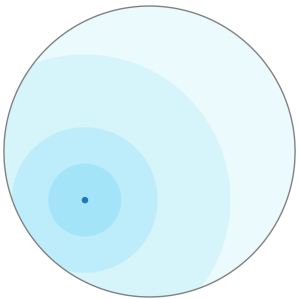
Definition (Beygelzimer et al. (2006))

A cover tree on $\mathcal{D}$ is a sequence $(\mathcal{C}_k)_k$ of coverings of $\mathcal{D}$:

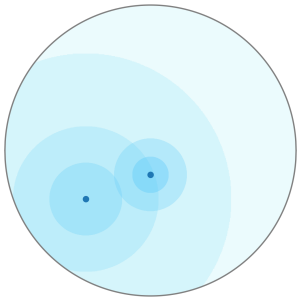
- *$\mathcal{C}_k$ is a subset of $\mathcal{D}$,*
- *$\mathcal{C}_k$ forms a 2^{-k}-net of $\mathcal{D}$,*
- *$\mathcal{C}_{k+1} \supset \mathcal{C}_k$.*



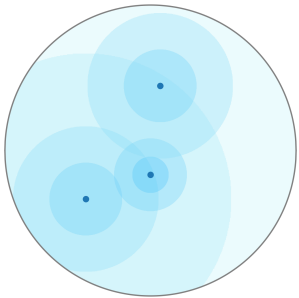
Sequential insertion for cover trees



$\text{root} \leftarrow X_1$



$\text{parent}(X_2) \leftarrow X_1$



$\text{parent}(X_3) \leftarrow X_1$

References

Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning*, pages 97–104, 2006.

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.

Tuomas Hytönen. A framework for non-homogeneous analysis on metric spaces, and the RBMO space of Tolsa. *Publicacions Matematiques*, pages 485–504, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, 24, 2011.