Online consistency of the nearest neighbor rule

Fall 2024 Seminar at Simons Institute

Sanjoy Dasgupta and Geelon So
October 23, 2024

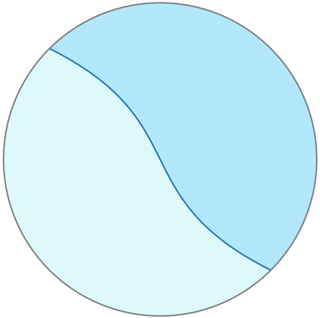
Outline of talk

1. Online classification
2. Some examples
3. Consistency on nice functions
4. Consistency on all functions
5. Broader ideas

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.



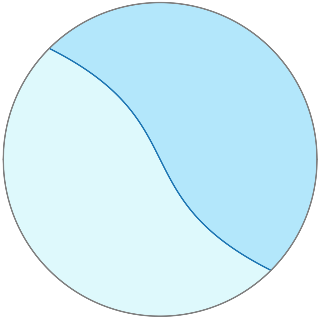
The realizable online setting

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.

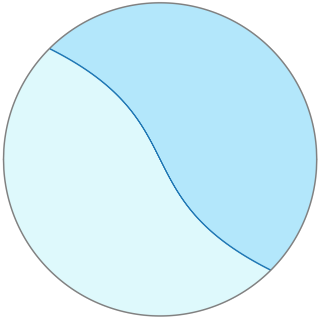


The realizable online setting

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.

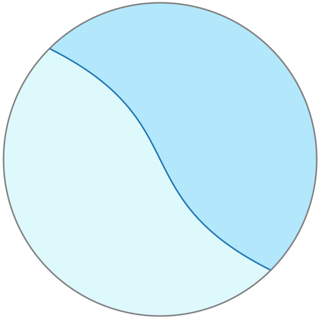
Online classification loop.
For $n = 1, 2, \ldots$

- 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.

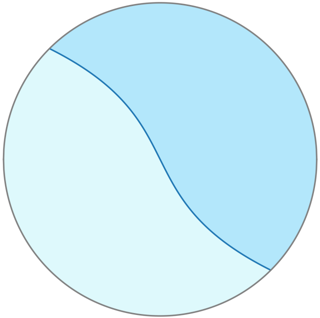


Online classification loop.
For $n = 1, 2, \ldots$

- 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

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

Nearest neighbor process

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

Definition

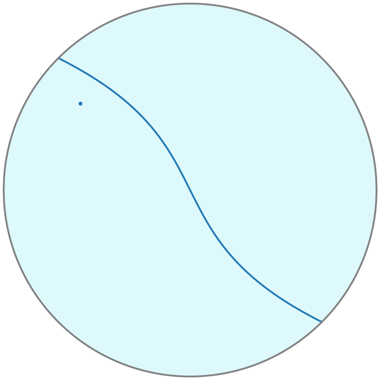
A nearest neighbor process is a sequence $\tilde{\mathbb{X}} = (\tilde{X}_n)_{n>0}$ satisfying

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

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

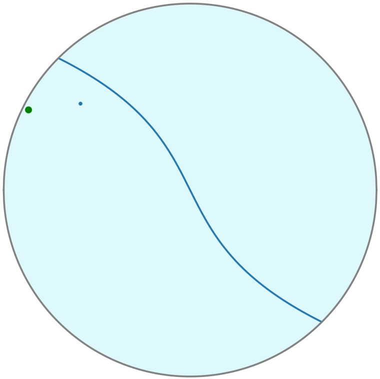
Behavior of the nearest neighbor rule in the i.i.d. setting.

I.I.D. sequence



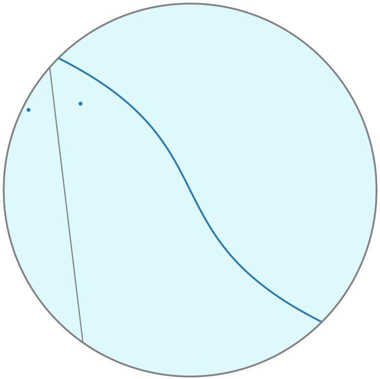
Time	0
Mistake counter	0

I.I.D. sequence



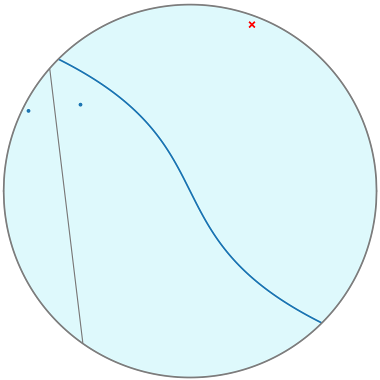
Time	1
Mistake counter	0

I.I.D. sequence



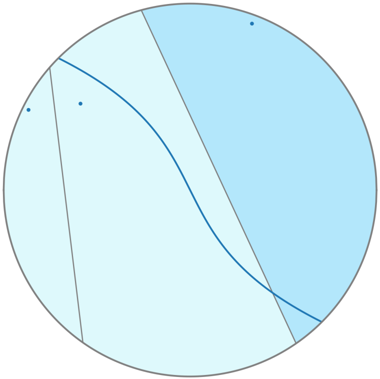
Time	1
Mistake counter	0

I.I.D. sequence



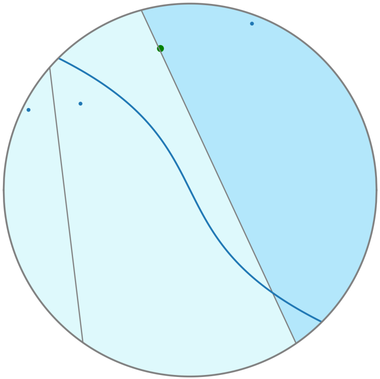
Time	2
Mistake counter	1

I.I.D. sequence



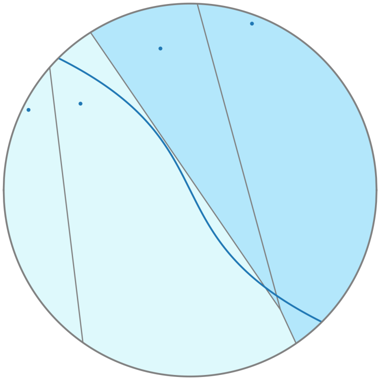
Time	2
Mistake counter	1

I.I.D. sequence



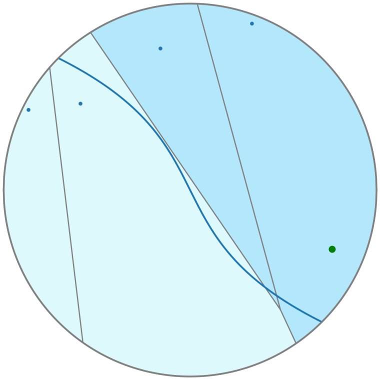
Time	3
Mistake counter	1

I.I.D. sequence



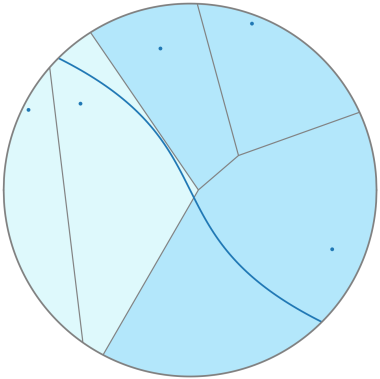
Time	3
Mistake counter	1

I.I.D. sequence



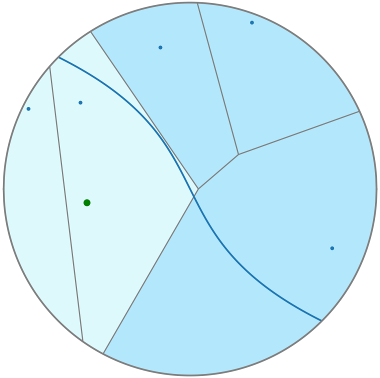
Time	4
Mistake counter	1

I.I.D. sequence



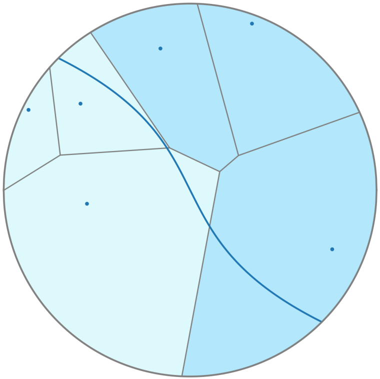
Time	4
Mistake counter	1

I.I.D. sequence



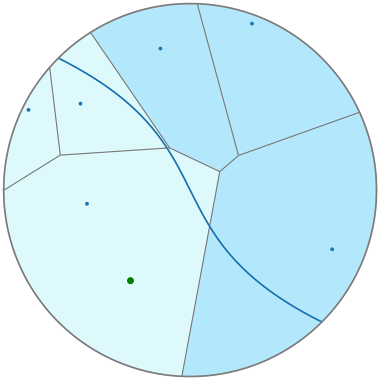
Time	5
Mistake counter	1

I.I.D. sequence



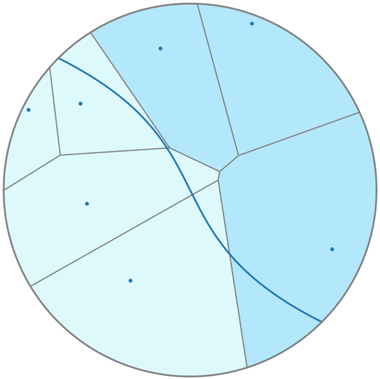
Time	5
Mistake counter	1

I.I.D. sequence



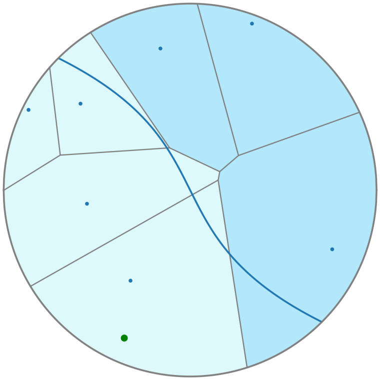
Time	6
Mistake counter	1

I.I.D. sequence



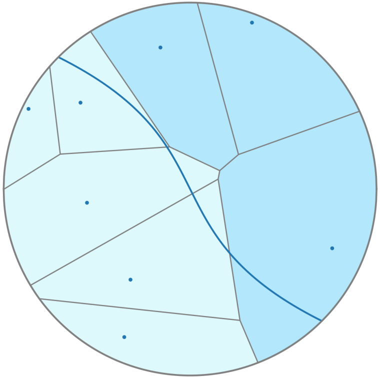
Time	6
Mistake counter	1

I.I.D. sequence



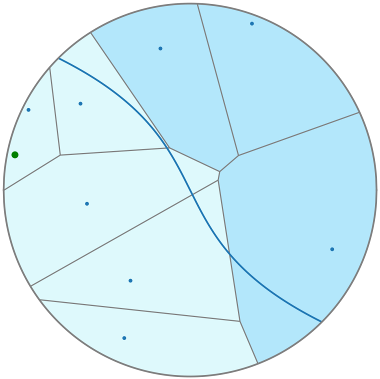
Time	7
Mistake counter	1

I.I.D. sequence



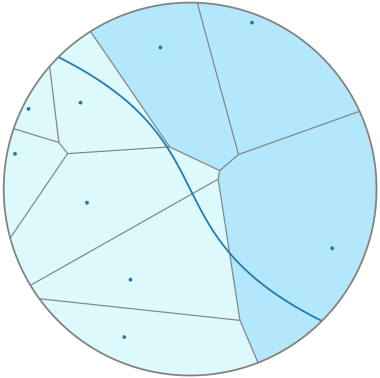
Time	7
Mistake counter	1

I.I.D. sequence



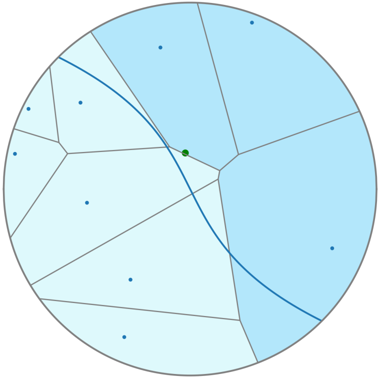
Time	8
Mistake counter	1

I.I.D. sequence



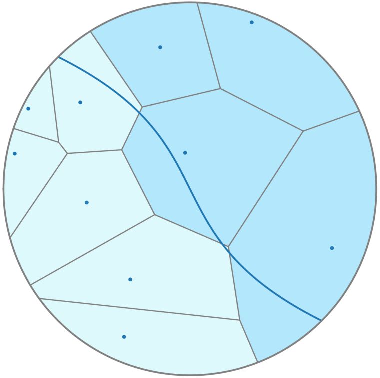
Time	8
Mistake counter	1

I.I.D. sequence



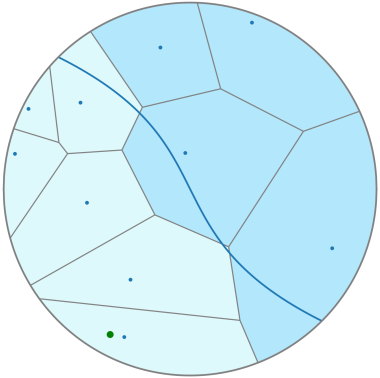
Time	9
Mistake counter	1

I.I.D. sequence



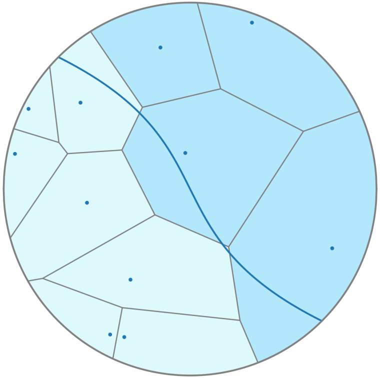
Time	9
Mistake counter	1

I.I.D. sequence



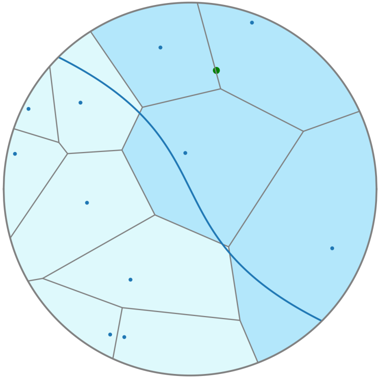
Time	10
Mistake counter	1

I.I.D. sequence



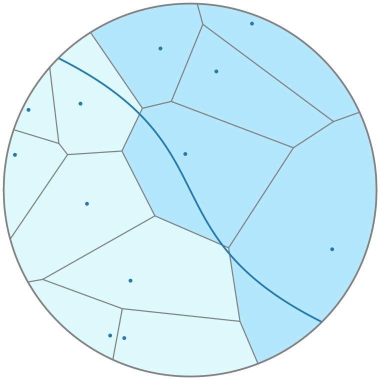
Time	10
Mistake counter	1

I.I.D. sequence



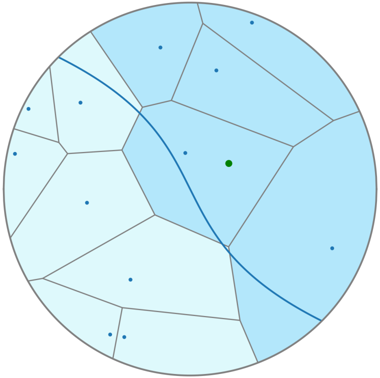
Time	11
Mistake counter	1

I.I.D. sequence



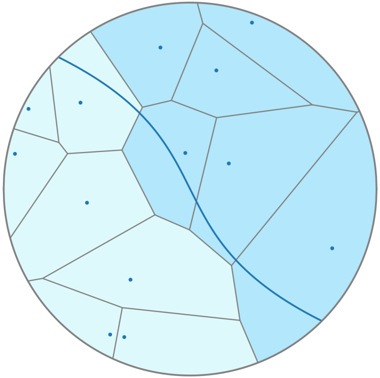
Time	11
Mistake counter	1

I.I.D. sequence



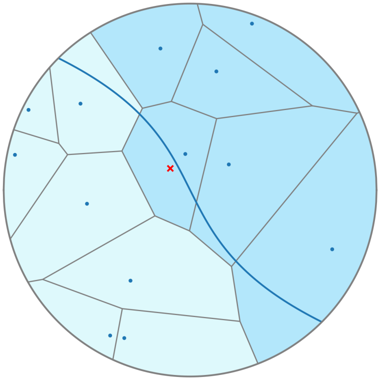
Time	12
Mistake counter	1

I.I.D. sequence



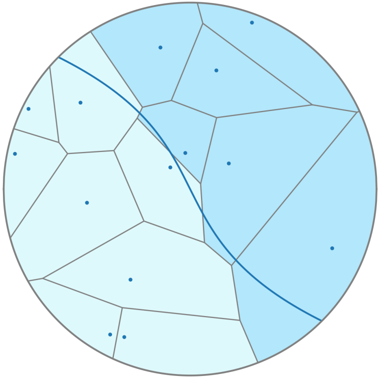
Time	12
Mistake counter	1

I.I.D. sequence



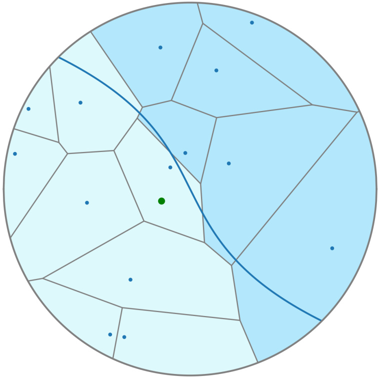
Time	13
Mistake counter	2

I.I.D. sequence



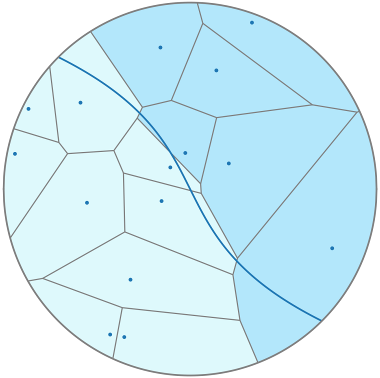
Time	13
Mistake counter	2

I.I.D. sequence



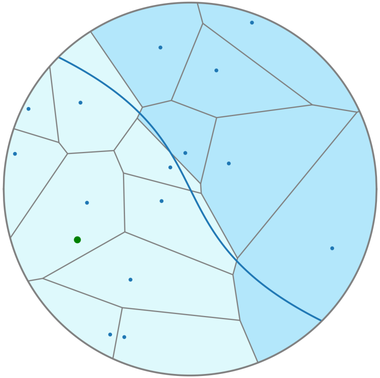
Time	14
Mistake counter	2

I.I.D. sequence



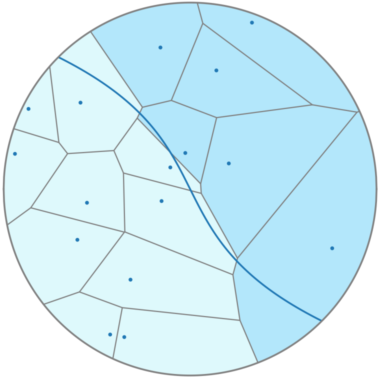
Time	14
Mistake counter	2

I.I.D. sequence



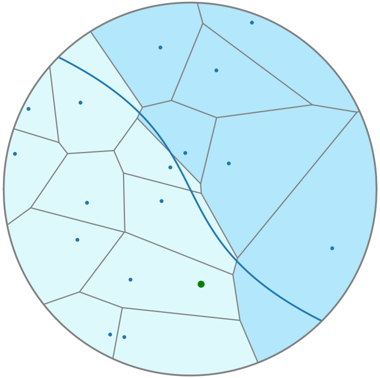
Time	15
Mistake counter	2

I.I.D. sequence



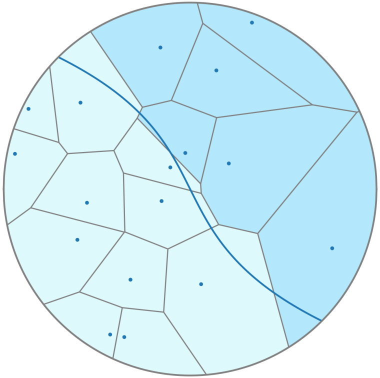
Time	15
Mistake counter	2

I.I.D. sequence



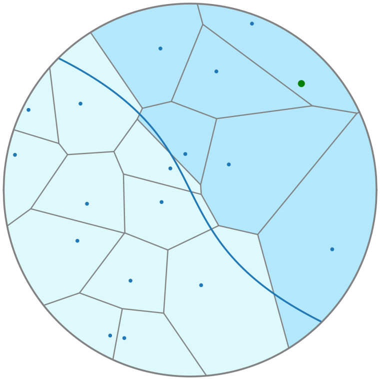
Time	16
Mistake counter	2

I.I.D. sequence



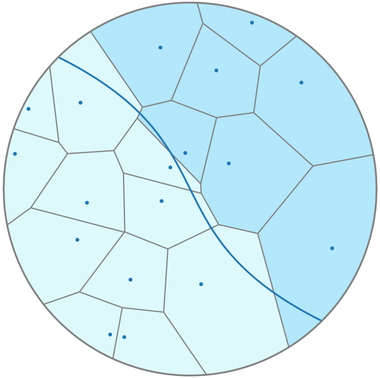
Time	16
Mistake counter	2

I.I.D. sequence



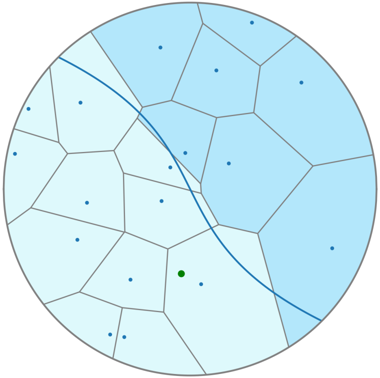
Time	17
Mistake counter	2

I.I.D. sequence



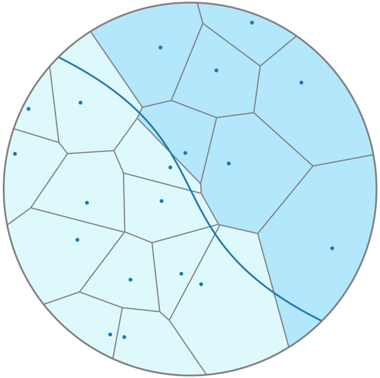
Time	17
Mistake counter	2

I.I.D. sequence



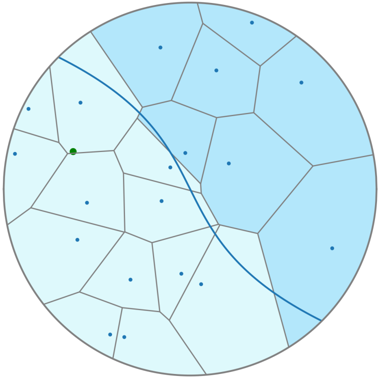
Time	18
Mistake counter	2

I.I.D. sequence



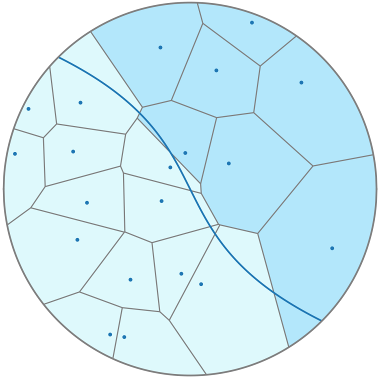
Time	18
Mistake counter	2

I.I.D. sequence



Time	19
Mistake counter	2

I.I.D. sequence

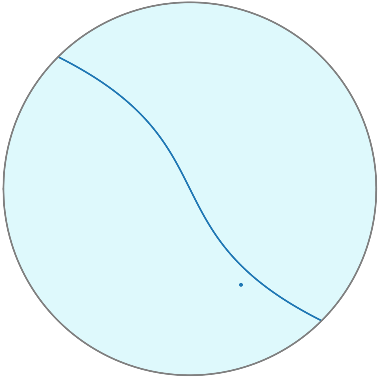


Time	19
Mistake counter	2

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

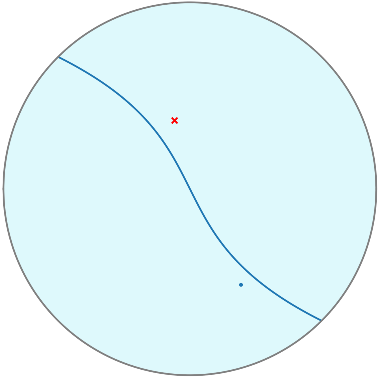
Worst-case sequence

Time	0
Mistake counter	0



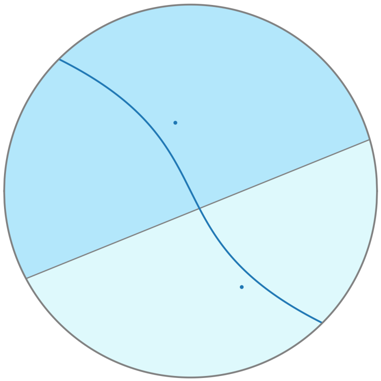
Worst-case sequence

Time	1
Mistake counter	1



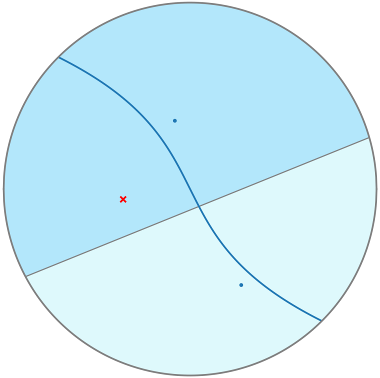
Worst-case sequence

Time	1
Mistake counter	1



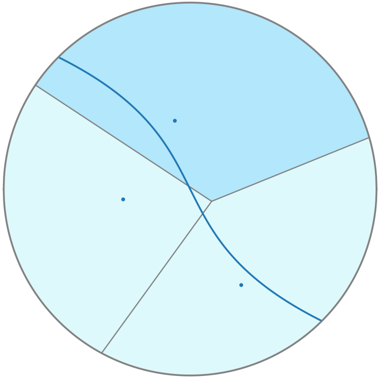
Worst-case sequence

Time	2
Mistake counter	2



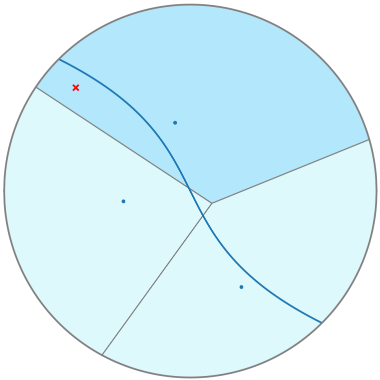
Worst-case sequence

Time	2
Mistake counter	2



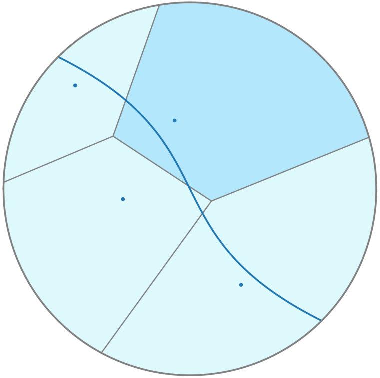
Worst-case sequence

Time	3
Mistake counter	3



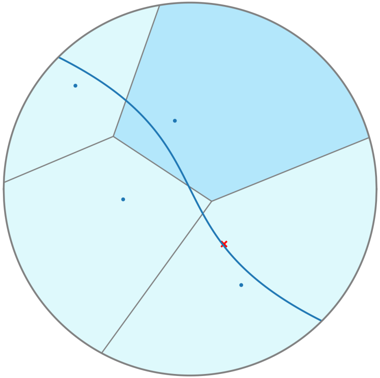
Worst-case sequence

Time	3
Mistake counter	3



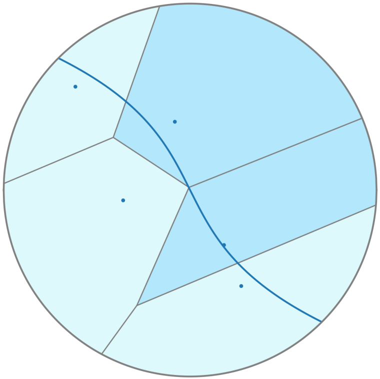
Worst-case sequence

Time	4
Mistake counter	4



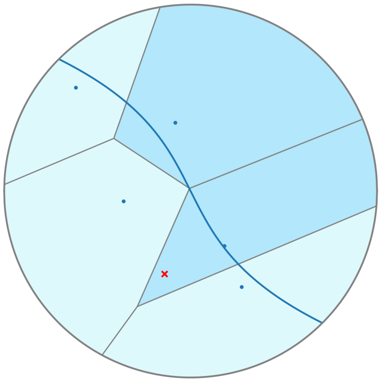
Worst-case sequence

Time	4
Mistake counter	4



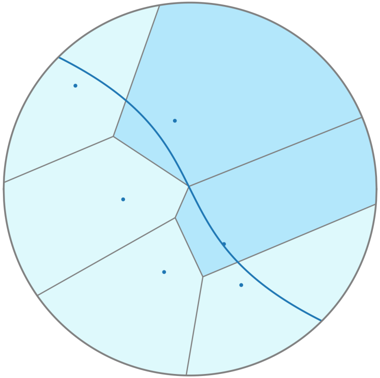
Worst-case sequence

Time	5
Mistake counter	5



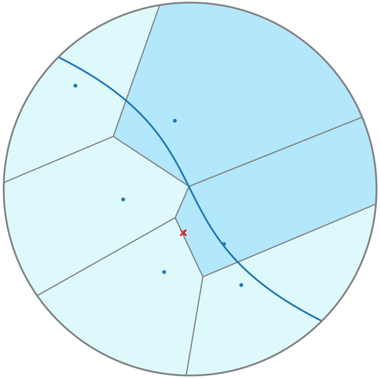
Worst-case sequence

Time	5
Mistake counter	5



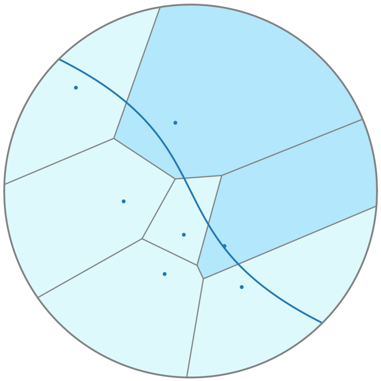
Worst-case sequence

Time	6
Mistake counter	6



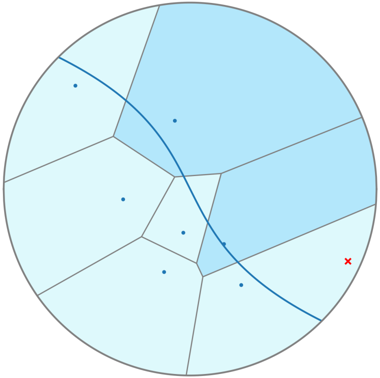
Worst-case sequence

Time	6
Mistake counter	6



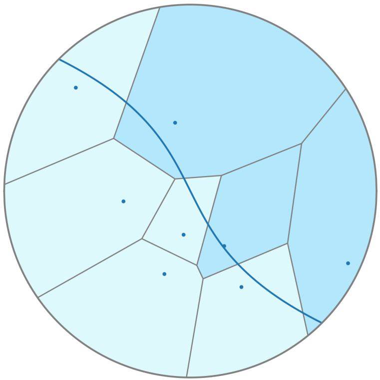
Worst-case sequence

Time	7
Mistake counter	7



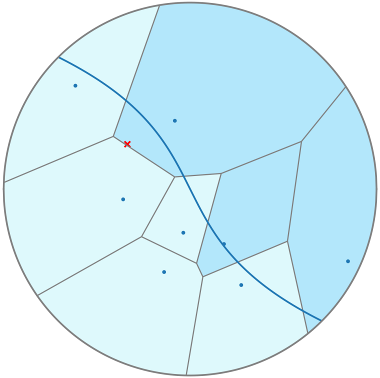
Worst-case sequence

Time	7
Mistake counter	7



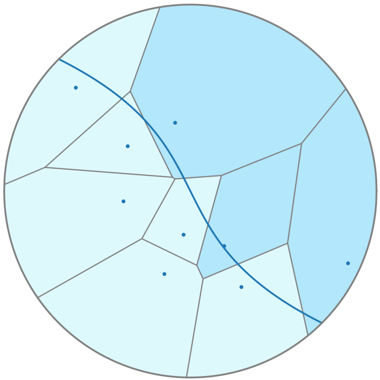
Worst-case sequence

Time	8
Mistake counter	8



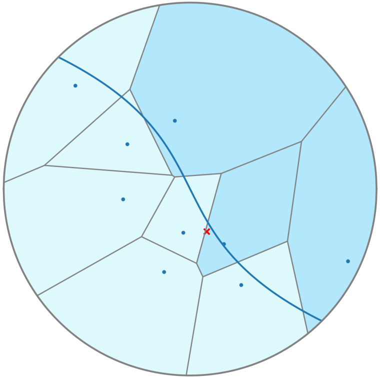
Worst-case sequence

Time	8
Mistake counter	8



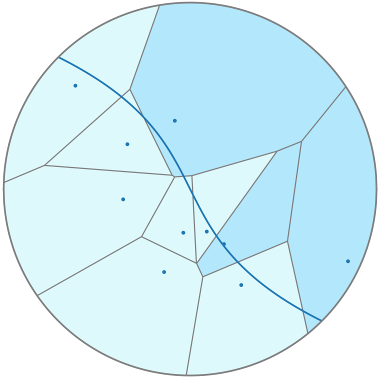
Worst-case sequence

Time	9
Mistake counter	9



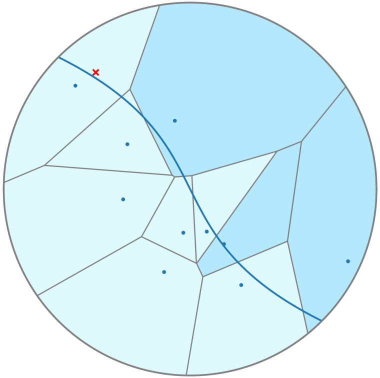
Worst-case sequence

Time	9
Mistake counter	9



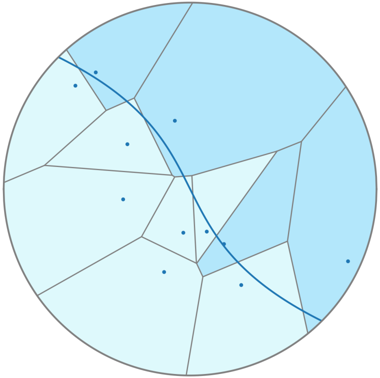
Worst-case sequence

Time	10
Mistake counter	10



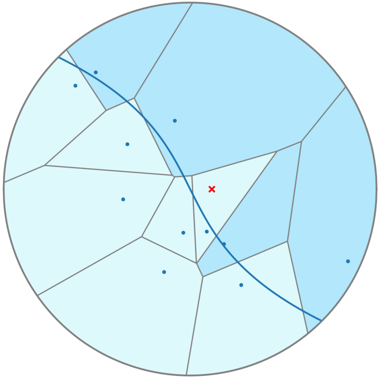
Worst-case sequence

Time	10
Mistake counter	10



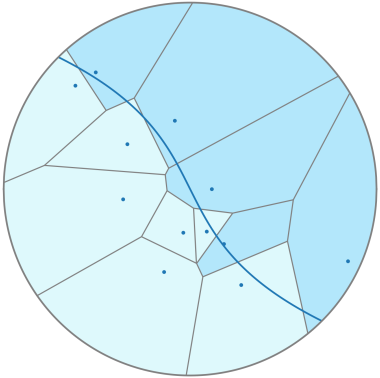
Worst-case sequence

Time	11
Mistake counter	11



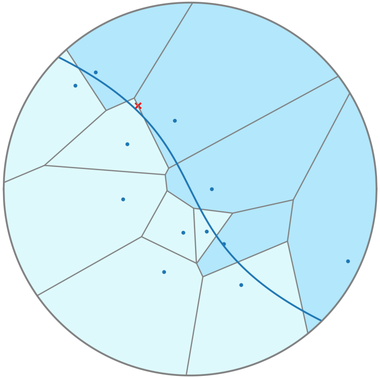
Worst-case sequence

Time	11
Mistake counter	11



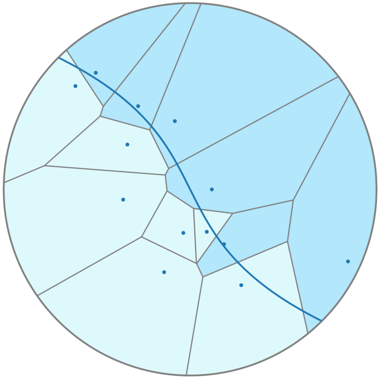
Worst-case sequence

Time	12
Mistake counter	12



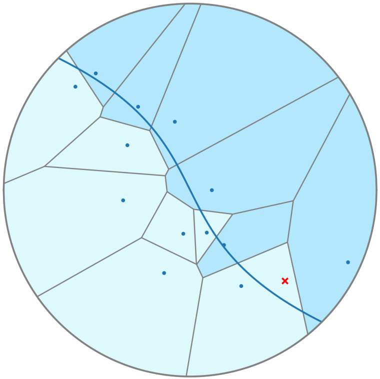
Worst-case sequence

Time	12
Mistake counter	12



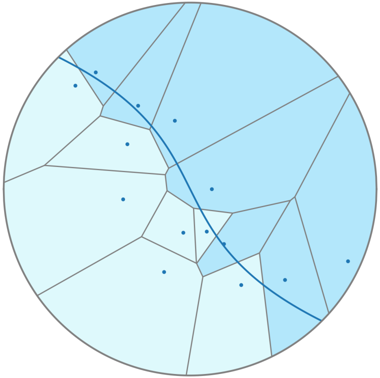
Worst-case sequence

Time	13
Mistake counter	13



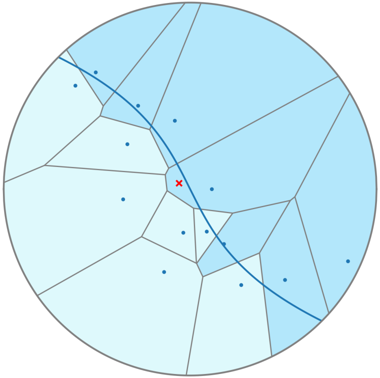
Worst-case sequence

Time	13
Mistake counter	13



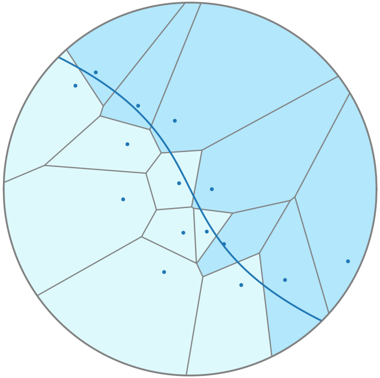
Worst-case sequence

Time	14
Mistake counter	14



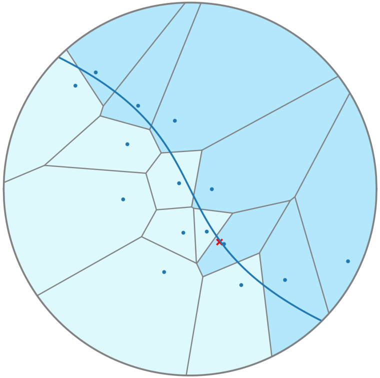
Worst-case sequence

Time	14
Mistake counter	14



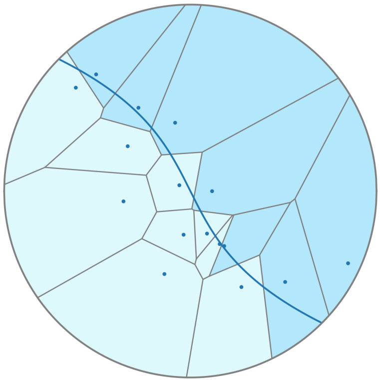
Worst-case sequence

Time	15
Mistake counter	15



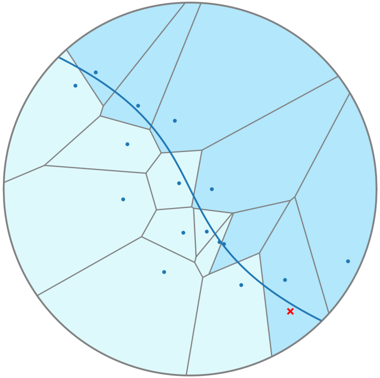
Worst-case sequence

Time	15
Mistake counter	15



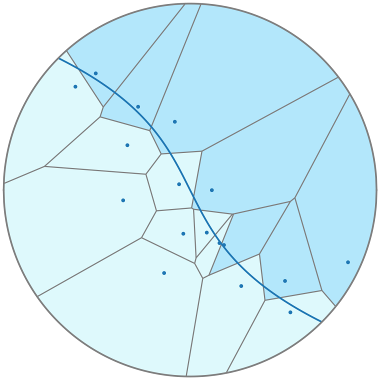
Worst-case sequence

Time	16
Mistake counter	16



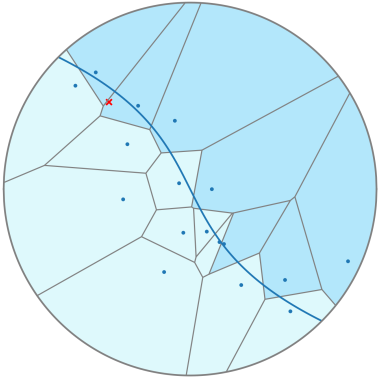
Worst-case sequence

Time	16
Mistake counter	16



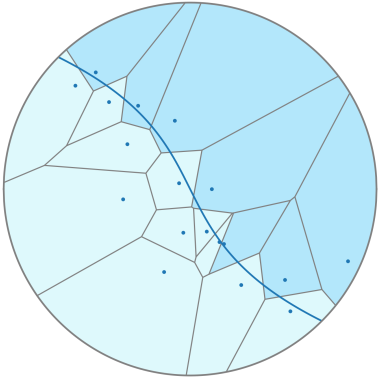
Worst-case sequence

Time	17
Mistake counter	17



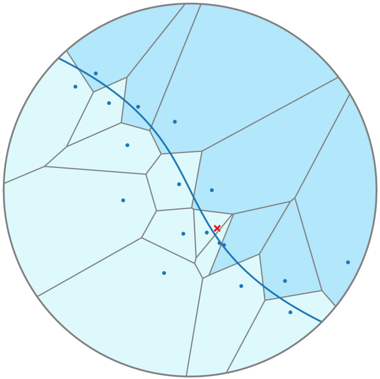
Worst-case sequence

Time	17
Mistake counter	17



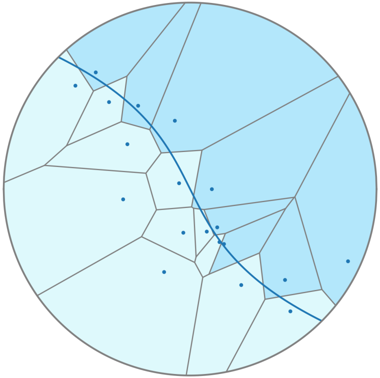
Worst-case sequence

Time	18
Mistake counter	18



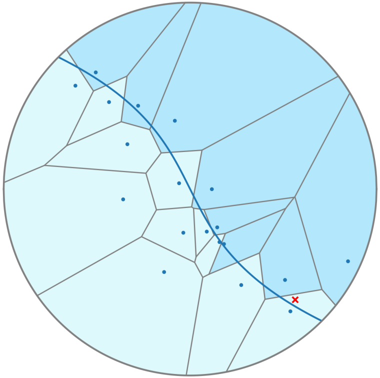
Worst-case sequence

Time	18
Mistake counter	18



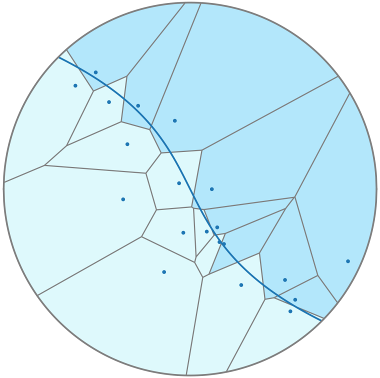
Worst-case sequence

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.

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:

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

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.

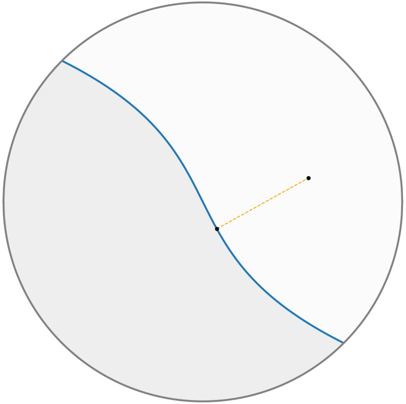
- Separable: every open cover has a countable subcover.
- Borel: we can measure the mass of balls.

Classification margin

Definition

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

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



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

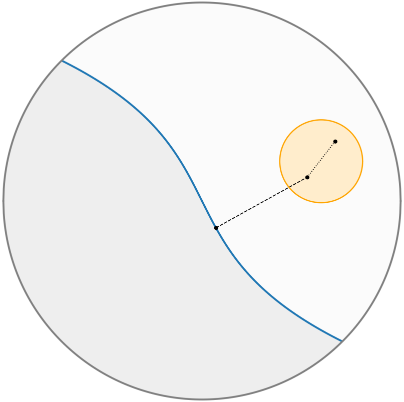
Definition

*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

*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:

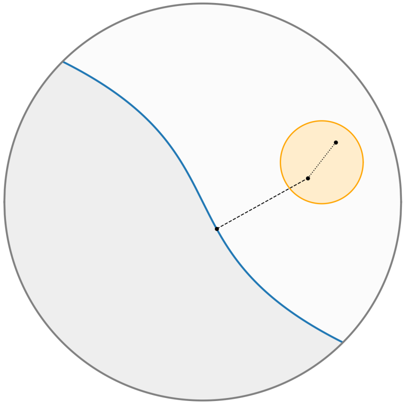
$$\operatorname{diam}(U) < \operatorname{margin}_\eta(x), \qquad \forall x \in U.$$

Proposition

Let x have positive margin:

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

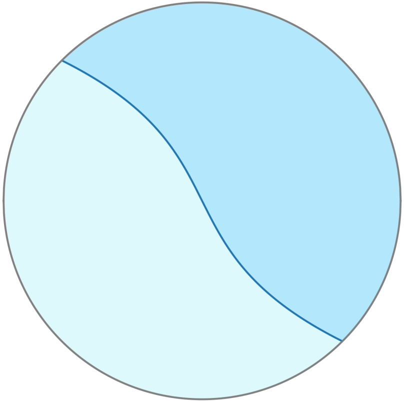
The open ball $B(x, r_x/3)$ is mutually labeling.



Functions with negligible boundaries

Definition

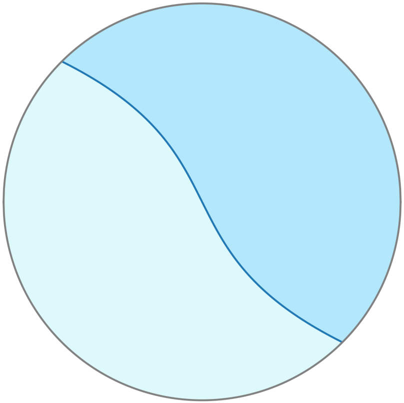
A function η *has* ***negligible boundary*** *if* ν*-almost all points have positive margin.*



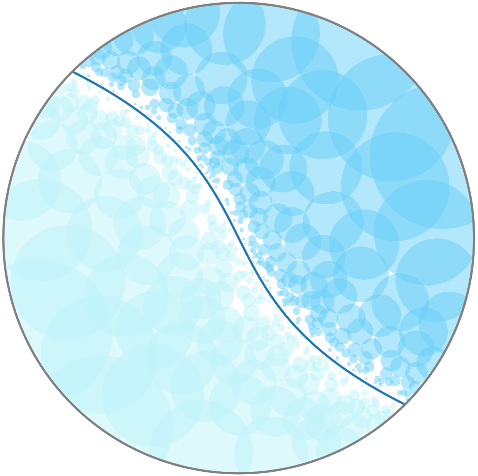
Functions with negligible boundaries

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 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.

↳ *since ρ is separable and ν is finite.*

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.

↳ *since ρ is separable and ν is finite.*

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

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:

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

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

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:*

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

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

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.

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:

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

Universal consistency on upper doubling spaces

Universal consistency

Goal: consistency for all measurable functions almost surely.

Universal consistency

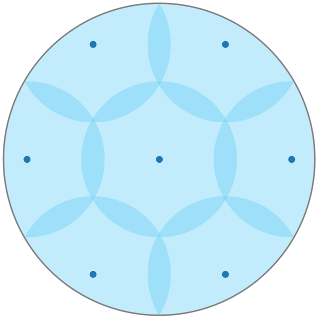
Goal: consistency for all measurable functions almost surely.

- 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.

- 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.

- Learning η is like learning η' when they have vanishingly small disagreement region $\{\eta \neq \eta'\}$.

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.

- 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:*

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

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

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*

Idea for universal consistency

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

Idea for universal consistency

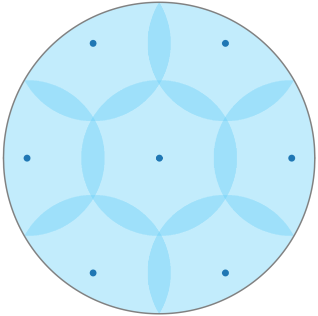
1. Even though 'bad points' can accumulate in memory, in a doubling space, their Voronoi cells tend to shrink (metric entropy) quickly as they are hit.
2. Suppose these Voronoi cells also shrink with respect to ν.
3. Then, it becomes increasingly unlikely that these bad points are nearest neighbors if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

Upper doubling measure

Definition

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

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



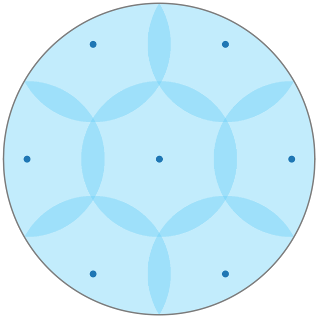
Upper doubling measure

Definition

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

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

Then, a small metric entropy implies small measure.



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.

Broader ideas

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 et al., 2021; Blanchard and Cosson, 2022; Blanchard, 2022)

Thank you!

Paper download: https://geelon.github.io

NeurIPS 2024

References

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.

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.

Steve Hanneke, Roi Livni, and Shay Moran. Online learning with simple predictors and a combinatorial characterization of minimax in 0/1 games. In *Conference on Learning Theory*, pages 2289–2314. PMLR, 2021.

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