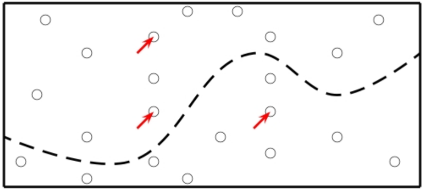
Adaptive Rates in Active Learning with Label Noise

Samory Kpotufe

Princeton University

Based on works with S. Ben David, R. Urner, A. Locatelli, A. Carpentier

Active Classification



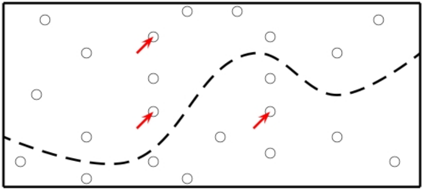
Pb: Classification $X \to Y \in \{0,1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

Applications:

Industrial: Document categorization, Vision/Audio, IoT security ...
Science: Medical imaging, Personalized medicine, Drug design ...

Q: Can active outperform passive learning? When? By how much?

Active Classification



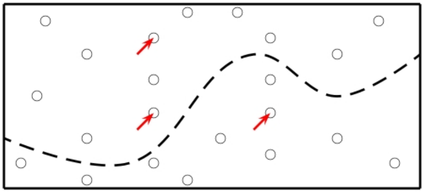
Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

Applications:

Industrial: Document categorization, Vision/Audio, IoT security ...
Science: Medical imaging, Personalized medicine, Drug design ...

Q: Can active outperform passive learning? When? By how much?

Active Classification



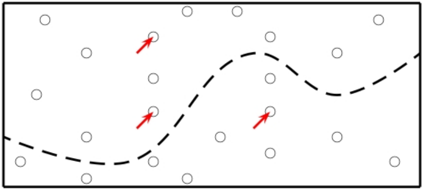
Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

Applications:

Industrial: Document categorization, Vision/Audio, IoT security ...
Science: Medical imaging, Personalized medicine, Drug design ...

Q: Can active outperform passive learning? When? By how much?

Active Classification



Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

Applications:

Industrial: Document categorization, Vision/Audio, IoT security ...
Science: Medical imaging, Personalized medicine, Drug design ...

Q: Can active outperform passive learning? When? By how much?

Gains in active learning

Performance measure:
- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in **parametric** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$: A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ is often $\gg 0$ (imperfect world):
noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in **parametric** *settings (e.g. VC dim.* $< \infty$*):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ **is often** $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in **parametric** *settings (e.g. VC dim.* $< \infty$*):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ **is often** $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ **is often** $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ **is often** $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ is often $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:
- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ is often $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ is often $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim.* $< \infty$*):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$**:** A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ **is often** $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Gains in active learning

Performance measure:

- Let f^* minimize $R(f) \doteq \mathbb{P}(Y \neq f(X))$.
- Let $\hat{f} \leftarrow$ classifier returned after querying n labels.

How small can $R(\hat{f}) - R(f^*)$ be in terms of n?

Most results are in ***parametric*** *settings (e.g. VC dim. $< \infty$):*

[Langford, Dasgupta, Hanneke, Balcan, et al ... since early 2000's]

$R(f^*) \approx 0$: A-L rates $\equiv e^{-\sqrt{n}}$, while P-L rates $\equiv 1/n$

$R(f^*) \gg 0$: A-L rates $\equiv 1/\sqrt{n}$ same as P-L rates.

But $R(f^*)$ is often $\gg 0$ (imperfect world):

noisy images or speech, adversarial spam, unpredictable drug response ...

Are there no gains in these practical settings?

Remarks:

Let $\eta(x) \doteq \mathbb{P}\left(Y = 1 \mid x\right)$, and note that $f^* = \mathbf{1}\left\{\eta \geq 1/2\right\}$.

So $R(f^*)$ depends on how η behaves.

A natural direction:

Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!

How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:

smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:
Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:
Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:
Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:
Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:
Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:
Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:

Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:

Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\Longrightarrow$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:
Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:
Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Remarks:

Let $\eta(x) \doteq \mathbb{P}(Y = 1 \mid x)$, and note that $f^* = \mathbf{1}\{\eta \geq 1/2\}$.
So $R(f^*)$ depends on how η behaves.

A natural direction:

Parametrize η on a **continuum** from **easy** to **hard** problems.

Capturing such continuum:

(i). Classification is hard if $\eta(x)$ is typically $\approx 1/2$, else it's easy!
How typical $\implies$ existing noise conditions (e.g. Tsyb., Mass., ...)

(ii). Combine with **regularity** or **complexity** conditions:
smoothness of η or class-boundary, complexity of hypothesis class ...

Initial insights in this direction ... different settings
[Hanneke 09], [Koltchinskii 10], [Castro-Nowak 08], [Minsker 12]

[Hanneke 09], [Koltchinskii 10] (ERM + low metric entropy):

Show considerable gains over passive learning even with label noise!

However:

- The above assume *bounded disagreement coefficient*:
  Mostly known for toy distributions ($\mathcal{U}$(interval), $\mathcal{U}$(sphere)).
- Procedures are not implementable (search over infinite $\mathcal{F}$).

What about **implementable** A-L procedures?

[Hanneke 09], [Koltchinskii 10] (ERM + low metric entropy):

Show considerable gains over passive learning even with label noise!

However:

- The above assume *bounded disagreement coefficient*:
  Mostly known for toy distributions ($\mathcal{U}$(interval), $\mathcal{U}$(sphere)).
- Procedures are not implementable (search over infinite $\mathcal{F}$).

What about **implementable** A-L procedures?

[Castro-Nowak 08] (smooth decision boundary):

Show considerable gains over passive learning even with label noise!

Implementable, no conditions on D-C!

However:

Needs full knowledge of boundary regularity and noise decay.

What about **adaptive + implementable** A-L procedures?

[Castro-Nowak 08] (smooth decision boundary):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C!

However:

Needs full knowledge of boundary regularity and noise decay.

What about **adaptive + implementable** A-L procedures?

[Castro-Nowak 08] (smooth decision boundary):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C!

However:

Needs full knowledge of boundary regularity and noise decay.

What about **adaptive + implementable** A-L procedures?

[Castro-Nowak 08] (smooth decision boundary):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C!

However:

Needs full knowledge of boundary regularity and noise decay.

What about **adaptive + implementable** A-L procedures?

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C, Adaptive!

However:
Needs quite restrictive technical conditions on $P_{X,Y}$.

What about **adaptive + implementable** A-L for **general** $P_{X,Y}$?

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C, Adaptive!

However:
Needs quite restrictive technical conditions on $P_{X,Y}$.

What about **adaptive + implementable** A-L for **general** $P_{X,Y}$?

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C, Adaptive!

However:
Needs quite restrictive technical conditions on $P_{X,Y}$.

What about **adaptive + implementable** A-L for **general** $P_{X,Y}$?

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on D-C, Adaptive!

However:
Needs quite restrictive technical conditions on $P_{X,Y}$.

What about **adaptive + implementable** A-L for **general** $P_{X,Y}$?

Outline:

We consider various regularity conditions on $\eta = \mathbb{E}[Y|X]$:

- η nearly aligns with clusters in X
  with R. Urner and S. Ben David, 2015
- η is a smooth function
  with A. Locatelli and A. Carpentier, 2017
- η defines a smooth decision-boundary
  with A. Locatelli and A. Carpentier, soon on Arxiv

Outline:

We consider various regularity conditions on $\eta = \mathbb{E}[Y|X]$:

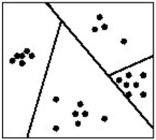
- η nearly aligns with clusters in X
  with R. Urner and S. Ben David, 2015
- η is a smooth function
  with A. Locatelli and A. Carpentier, 2017
- η defines a smooth decision-boundary
  with A. Locatelli and A. Carpentier, soon on Arxiv

η *nearly aligns with clusters in* X

Related to the *cluster assumption* (C-A):

One label dominates in each cluster

So query $O(1)$ labels per cluster



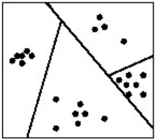
Benefits: Few label queries when C-A holds! Implementable!

Downside: unsafe assumption!

Fortunately there are existing safe approaches ...

η *nearly aligns with clusters in* X

Related to the *cluster assumption* (C-A):
One label dominates in each cluster
So query $O(1)$ labels per cluster



Benefits: Few label queries when C-A holds! Implementable!
Downside: unsafe assumption!

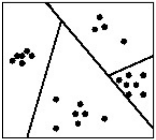
Fortunately there are existing safe approaches ...

η *nearly aligns with clusters in* X

Related to the *cluster assumption* (C-A):

One label dominates in each cluster

So query $O(1)$ labels per cluster



Benefits: Few label queries when C-A holds! Implementable!

Downside: unsafe assumption!

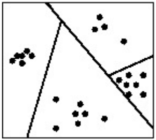
Fortunately there are existing safe approaches ...

η *nearly aligns with clusters in* X

Related to the *cluster assumption* (C-A):

One label dominates in each cluster

So query $O(1)$ labels per cluster



Benefits: Few label queries when C-A holds! Implementable!

Downside: unsafe assumption!

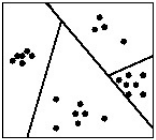
Fortunately there are existing safe approaches ...

η *nearly aligns with clusters in* X

Related to the *cluster assumption* (C-A):

One label dominates in each cluster

So query $O(1)$ labels per cluster

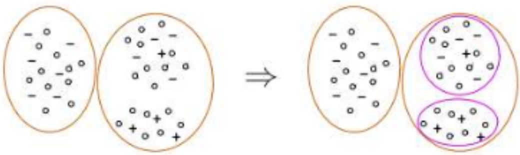


Benefits: Few label queries when C-A holds! Implementable!

Downside: unsafe assumption!

Fortunately there are existing safe approaches ...

Hierarchical Labeling: Dasgupta and Hsu 2008



− Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

• If there is a clear majority label (say $1 - \epsilon$ proportion):
    LABEL the cell (using majority label)

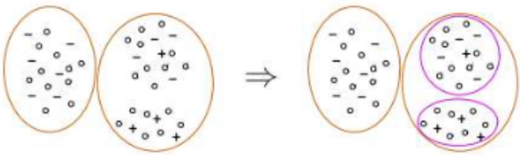
• Else, PARTITION the cell and REPEAT

Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X
C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: *Dasgupta and Hsu 2008*



– Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

• If there is a clear majority label (say $1 - \epsilon$ proportion):
LABEL the cell (using majority label)

• Else, PARTITION the cell and REPEAT

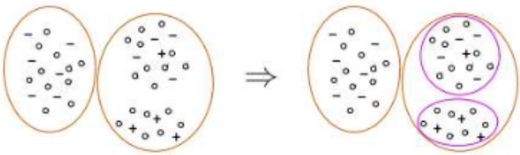
Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X

C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: Dasgupta and Hsu 2008



$-$ Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

• If there is a clear majority label (say $1 - \epsilon$ proportion):
  LABEL the cell (using majority label)

• Else, PARTITION the cell and REPEAT

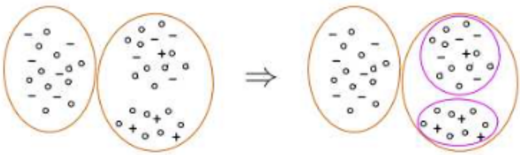
Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X

C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: *Dasgupta and Hsu 2008*



$-$ Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

- If there is a clear majority label (say $1-\epsilon$ proportion):
  LABEL the cell (using majority label)
- Else, PARTITION the cell and REPEAT

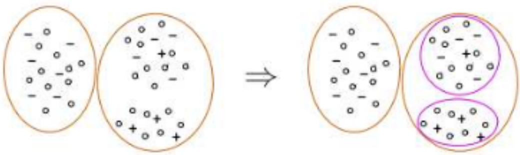
Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X

C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: *Dasgupta and Hsu 2008*



$-$ Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

- If there is a clear majority label (say $1 - \epsilon$ proportion):
  LABEL the cell (using majority label)
- Else, PARTITION the cell and REPEAT

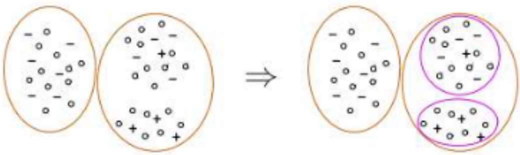
Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X

C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: *Dasgupta and Hsu 2008*



$-$ Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

- If there is a clear majority label (say $1-\epsilon$ proportion):
  LABEL the cell (using majority label)
- Else, PARTITION the cell and REPEAT

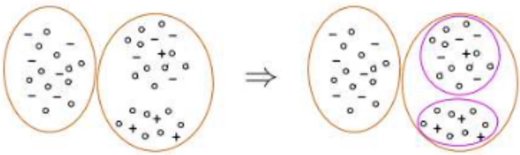
Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X

C-A: savings when C-A nearly holds, **SAFE** when not.

Hierarchical Labeling: *Dasgupta and Hsu 2008*



− Partition unlabeled X_1^n, query a few labels in each cell.

Consider each cell:

- If there is a clear majority label (say $1 - \epsilon$ proportion):
  LABEL the cell (using majority label)
- Else, PARTITION the cell and REPEAT

Label data with error $< \epsilon \implies$ now use supervised learner.

Overall Appeal:

A-L: Implementable and has guarantees for general P_X
C-A: savings when C-A nearly holds, **SAFE** when not.

Labeling Goal: $\le 1/\epsilon^2$ label-queries of agnostic-learning.

Guarantees on label-queries: from $|T|_* \cdot (1/\epsilon)$ to $1/\epsilon^2$

Depends on niceness of $P_{X,Y}$, and $|T|_* \equiv$ Data-quantization rate.

Earlier results (similar label guarantees)

- [Das., Hsu, 08]: Niceness of sample X_1^n, Y_1^n.
- [Urn., Wulff, B-Dav, 13]: Niceness of $P_{X,Y}$, **no noise in** Y, partition T **cannot depend on** X_1^n.

Our results: *more practical assumptions*

Niceness of $P_{X,Y}$, low noise in Y, $T = T(X_1^n) \implies$ smaller $|T|_*$.

We now have $|T|_* = O(2^d) \ll 2^D$, ($d =$ intrinsic dim. of X).

Labeling Goal: $\leq 1/\epsilon^2$ label-queries of agnostic-learning.

Guarantees on label-queries: from $|T|_* \cdot (1/\epsilon)$ to $1/\epsilon^2$

Depends on niceness of $P_{X,Y}$, and $|T|_* \equiv$ Data-quantization rate.

Earlier results (similar label guarantees)

- [Das., Hsu, 08]: Niceness of sample X_1^n, Y_1^n.
- [Urn., Wulff, B-Dav, 13]: Niceness of $P_{X,Y}$, **no noise in** Y, partition T **cannot depend on** X_1^n.

Our results: *more practical assumptions*

Niceness of $P_{X,Y}$, low noise in Y, $T = T(X_1^n) \implies$ smaller $|T|_*$.

We now have $|T|_* = O(2^d) \ll 2^D$, ($d =$ intrinsic dim. of X).

Labeling Goal: $\leq 1/\epsilon^2$ label-queries of agnostic-learning.

Guarantees on label-queries: from $|T|_* \cdot (1/\epsilon)$ to $1/\epsilon^2$

Depends on niceness of $P_{X,Y}$, and $|T|_* \equiv$ Data-quantization rate.

Earlier results (similar label guarantees)

- [Das., Hsu, 08]: Niceness of sample X_1^n, Y_1^n.
- [Urn., Wulff, B-Dav, 13]: Niceness of $P_{X,Y}$, **no noise in** Y, partition T **cannot depend on** X_1^n.

Our results: *more practical assumptions*

Niceness of $P_{X,Y}$, low noise in Y, $T = T(X_1^n) \implies$ smaller $|T|_*$.

We now have $|T|_* = O(2^d) \ll 2^D$, ($d =$ intrinsic dim. of X).

Labeling Goal: $\leq 1/\epsilon^2$ label-queries of agnostic-learning.

Guarantees on label-queries: from $|T|_* \cdot (1/\epsilon)$ to $1/\epsilon^2$

Depends on niceness of $P_{X,Y}$, and $|T|_* \equiv$ Data-quantization rate.

Earlier results (similar label guarantees)

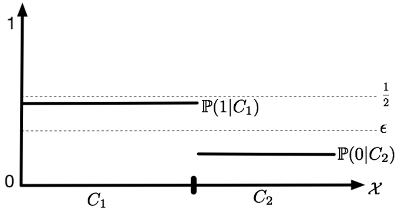
- [Das., Hsu, 08]: Niceness of sample X_1^n, Y_1^n.
- [Urn., Wulff, B-Dav, 13]: Niceness of $P_{X,Y}$, **no noise in** Y, partition T **cannot depend on** X_1^n.

Our results: *more practical assumptions*

Niceness of $P_{X,Y}$, low noise in Y, $T = T(X_1^n) \implies$ smaller $|T|_*$.

We now have $|T|_* = O(2^d) \ll 2^D$, ($d =$ intrinsic dim. of X).

Niceness (or parametrization) of $P_{X,Y}$



Two main conditions on $\eta(x) = \mathbb{E}[Y|x]$:

η is likely far from $\frac{1}{2}$ (Tsy. noise condition):

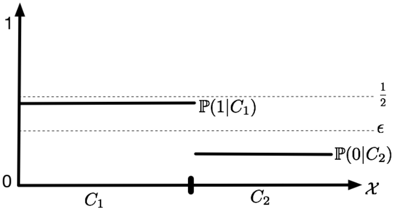
$$\mathbb{P}_X(|\eta(X) - 1/2| < \tau) \leq \tau^\beta$$

η is nearly Lipschitz:

$$\mathbb{P}_X(\exists x \text{ s.t. } |\eta(X) - \eta(x)| > \lambda \|X - x\|) \leq \lambda^{-\alpha}$$

Large $\alpha, \beta \implies$ **C-A holds** (at least for small clusters).

Niceness (or parametrization) of $P_{X,Y}$



Two main conditions on $\eta(x) = \mathbb{E}\left[Y|x\right]$:

η is likely far from $\frac{1}{2}$ (Tsy. noise condition):

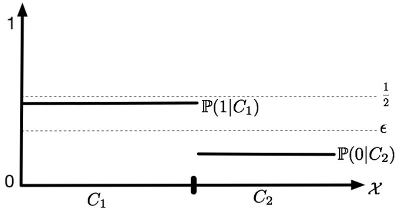
$$\mathbb{P}_X\left(|\eta(X) - 1/2| < \tau\right) \leq \tau^{\beta}$$

η is nearly Lipschitz:

$$\mathbb{P}_X\left(\exists x \text{ s.t. } |\eta(X) - \eta(x)| > \lambda\|X - x\|\right) \leq \lambda^{-\alpha}$$

Large $\alpha, \beta \implies$ **C-A holds** (at least for small clusters).

Niceness (or parametrization) of $P_{X,Y}$



Two main conditions on $\eta(x) = \mathbb{E}[Y|x]$:

η is likely far from $\frac{1}{2}$ (Tsy. noise condition):

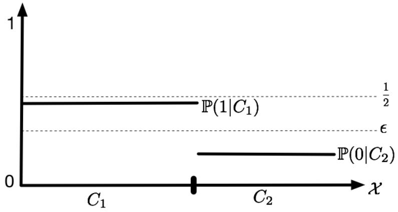
$$\mathbb{P}_X\left(|\eta(X) - 1/2| < \tau\right) \leq \tau^{\beta}$$

η is nearly Lipschitz:

$$\mathbb{P}_X\left(\exists\, x \text{ s.t. } |\eta(X) - \eta(x)| > \lambda\|X - x\|\right) \leq \lambda^{-\alpha}$$

Large $\alpha, \beta \implies$ **C-A holds** (at least for small clusters).

Niceness (or parametrization) of $P_{X,Y}$



Two main conditions on $\eta(x) = \mathbb{E}[Y|x]$:

η is likely far from $\frac{1}{2}$ (Tsy. noise condition):

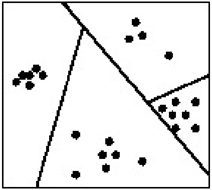
$$\mathbb{P}_X(|\eta(X) - 1/2| < \tau) \leq \tau^\beta$$

η is nearly Lipschitz:

$$\mathbb{P}_X(\exists\, x \text{ s.t. } |\eta(X) - \eta(x)| > \lambda\|X - x\|) \leq \lambda^{-\alpha}$$

Large $\alpha, \beta \implies$ **C-A holds** (at least for small clusters).

Parametrizing the partition-tree T



Two main ingredients:

Cells of T have bounded complexity V_T

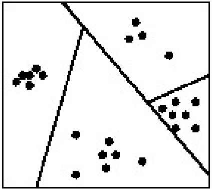
Allows for decoupling the dependence between $T(X_1^n)$ and X_1^n.

T has good quantization rate

Let $T_r \equiv$ level where cells have diameter r; $|T_r| \lesssim r^{-\kappa}$

Remark: $\kappa = O(d) \ll D$ w.h.p. for various procedures T (Rand. Proj., PCA, Rand. k-d) [Verma, Kpo., Das. 10] [Vemp. 12]

Parametrizing the partition-tree T



Two main ingredients:

Cells of T have bounded complexity V_T

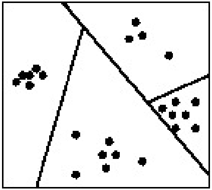
Allows for decoupling the dependence between $T(X_1^n)$ and X_1^n.

T has good quantization rate

Let $T_r \equiv$ level where cells have diameter r; $|T_r| \lesssim r^{-\kappa}$

Remark: $\kappa = O(d) \ll D$ w.h.p. for various procedures T (Rand. Proj., PCA, Rand. k-d) [Verma, Kpo., Das. 10] [Vemp. 12]

Parametrizing the partition-tree T



Two main ingredients:

Cells of T have bounded complexity V_T

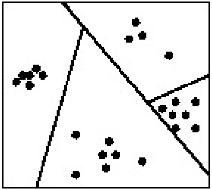
Allows for decoupling the dependence between $T(X_1^n)$ and X_1^n.

T has good quantization rate

Let $T_r \equiv$ level where cells have diameter r; $|T_r| \lesssim r^{-\kappa}$

Remark: $\kappa = O(d) \ll D$ w.h.p. for various procedures T (Rand. Proj., PCA, Rand. k-d) [Verma, Kpo., Das. 10] [Vemp. 12]

Parametrizing the partition-tree T



Two main ingredients:

Cells of T have bounded complexity V_T

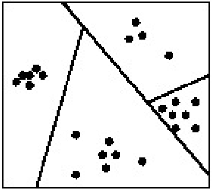
Allows for decoupling the dependence between $T(X_1^n)$ and X_1^n.

T has good quantization rate

Let $T_r \equiv$ level where cells have diameter r; $|T_r| \lesssim r^{-\kappa}$

Remark: $\kappa = O(d) \ll D$ w.h.p. for various procedures T (Rand. Proj., PCA, Rand. k-d) [Verma, Kpo., Das. 10] [Vemp. 12]

Parametrizing the partition-tree T



Two main ingredients:

Cells of T have bounded complexity V_T

Allows for decoupling the dependence between $T(X_1^n)$ and X_1^n.

T has good quantization rate

Let $T_r \equiv$ level where cells have diameter r; $|T_r| \lesssim r^{-\kappa}$

Remark: $\kappa = O(d) \ll D$ w.h.p. for various procedures T
(Rand. Proj., PCA, Rand. k-d) [Verma, Kpo., Das. 10] [Vemp. 12]

Guarantees (w.h.p.)

Given: $n = \Omega(1/\epsilon^2)$ unlabeled samples X_1^n.

- **Correctness:** At most ϵ fraction of X_1^n is *mislabeled*.
- **Labels requested:** At most

$$n \cdot \left(2^{\kappa/(1+\kappa/\alpha)} \cdot \epsilon^{1/(1+\kappa/\alpha)} + \exp(-\epsilon \cdot \beta)\right)$$

- This is best as C-A holds (α, β large), safe if not.
- Avoids the curse of dimension for structured data ($\kappa \approx d \ll D$).

Guarantees (w.h.p.)

Given: $n = \Omega(1/\epsilon^2)$ unlabeled samples X_1^n.

- **Correctness:** At most ϵ fraction of X_1^n is *mislabeled*.
- **Labels requested:** At most

$$n \cdot \left(2^{\kappa/(1+\kappa/\alpha)} \cdot \epsilon^{1/(1+\kappa/\alpha)} + \exp(-\epsilon \cdot \beta)\right)$$

- This is best as C-A holds (α, β large), safe if not.
- Avoids the curse of dimension for structured data ($\kappa \approx d \ll D$).

Guarantees (w.h.p.)

Given: $n = \Omega(1/\epsilon^2)$ unlabeled samples X_1^n.

- **Correctness:** At most ϵ fraction of X_1^n is *mislabeled*.
- **Labels requested:** At most

$$n \cdot \left(2^{\kappa/(1+\kappa/\alpha)} \cdot \epsilon^{1/(1+\kappa/\alpha)} + \exp(-\epsilon \cdot \beta)\right)$$

- This is best as C-A holds (α, β large), safe if not.
- Avoids the curse of dimension for structured data ($\kappa \approx d \ll D$).

Guarantees (w.h.p.)

Given: $n = \Omega(1/\epsilon^2)$ unlabeled samples X_1^n.

- **Correctness:** At most ϵ fraction of X_1^n is *mislabeled*.
- **Labels requested:** At most

$$n \cdot \left(2^{\kappa/(1+\kappa/\alpha)} \cdot \epsilon^{1/(1+\kappa/\alpha)} + \exp(-\epsilon \cdot \beta)\right)$$

- This is best as C-A holds (α, β large), safe if not.

- Avoids the curse of dimension for structured data ($\kappa \approx d \ll D$).

Guarantees (w.h.p.)

Given: $n = \Omega(1/\epsilon^2)$ unlabeled samples X_1^n.

- **Correctness:** At most ϵ fraction of X_1^n is *mislabeled*.
- **Labels requested:** At most

$$n \cdot \left(2^{\kappa/(1+\kappa/\alpha)} \cdot \epsilon^{1/(1+\kappa/\alpha)} + \exp(-\epsilon \cdot \beta) \right)$$

- This is best as C-A holds (α, β large), safe if not.
- Avoids the curse of dimension for structured data ($\kappa \approx d \ll D$).

Outline:

We consider various regularity conditions on $\eta = \mathbb{E}[Y|X]$:

- η nearly aligns with clusters in X
  with R. Urner and S. Ben David, 2015
- η is a smooth function
  with A. Locatelli and A. Carpentier, 2017
- η defines a smooth decision-boundary
  with A. Locatelli and A. Carpentier, soon on Arxiv

η *is a smooth function*

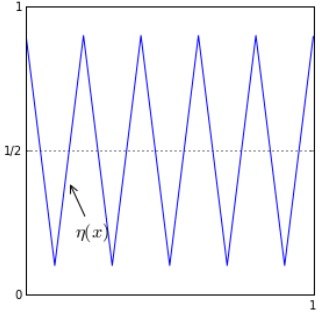
Setup:

- $\eta(x) \doteq \mathbb{E}[Y|x]$ has Hölder smoothness α
  (e.g. all derivatives up to order α are bounded)
- Tsybakov noise condition: $\exists c, \beta \geq 0$ such that $\forall \tau > 0$:

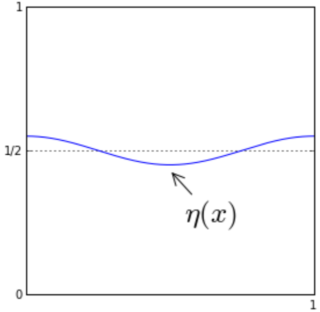
$$\mathbb{P}_X\left(x : \left|\eta(x) - \frac{1}{2}\right| \leq \tau\right) \leq c\tau^{\beta},$$

…

α and β: continuum between easy and hard problems



Small α

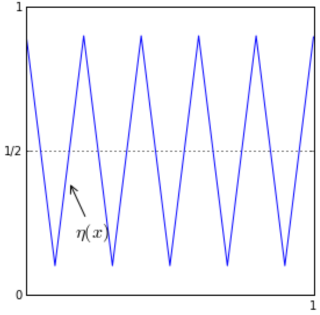


Small β

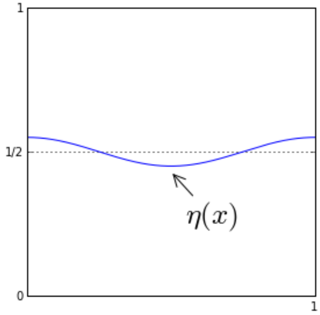
Questions: how do α, β and d interact? Can we adapt to this?

...

α and β: continuum between easy and hard problems



Small α



Small β

Questions: how do α, β and d interact? Can we adapt to this?

Previous work Minsker (2012): $\mathbb{P}_X$ uniform

Self-similarity of η: smoothness is tight $\forall x$ (never better than α)

Theorem: $\alpha \leq 1$, $\alpha\beta \leq d$

There exists an active strategy $\hat{f}_n$ such that:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}} \quad \text{(rate is tight)}$$

Passive rate: replace $d - \alpha\beta$ by d [AT07]

For $\alpha > 1$ the rate seems to transition:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

Minsker conjectures that this rate is tight.

Open: Unrestricted $\mathbb{P}_X$? General η? Tightness of $\alpha > 1$?

Previous work Minsker (2012): $\mathbb{P}_X$ uniform

Self-similarity of η: smoothness is tight $\forall x$ (never better than α)

Theorem: $\alpha \leq 1$, $\alpha\beta \leq d$

There exists an active strategy $\hat{f}_n$ such that:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}} \quad \text{(rate is tight)}$$

Passive rate: replace $d - \alpha\beta$ by d [AT07]

For $\alpha > 1$ the rate seems to transition:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

Minsker conjectures that this rate is tight.

Open: Unrestricted $\mathbb{P}_X$? General η? Tightness of $\alpha > 1$?

Previous work Minsker (2012): $\mathbb{P}_X$ uniform

Self-similarity of η: smoothness is tight $\forall x$ (never better than α)

Theorem: $\alpha \leq 1$, $\alpha\beta \leq d$

There exists an active strategy $\hat{f}_n$ such that:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}} \quad \text{(rate is tight)}$$

Passive rate: replace $d - \alpha\beta$ by d [AT07]

For $\alpha > 1$ the rate seems to transition:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

Minsker conjectures that this rate is tight.

Open: Unrestricted $\mathbb{P}_X$? General η? Tightness of $\alpha > 1$?

Our results: statistical contributions

Milder conditions, new rate regimes

- $\mathbb{P}_X$ uniform: same rates without self-similarity condition
- Verify rate transition for $\alpha > 1$:

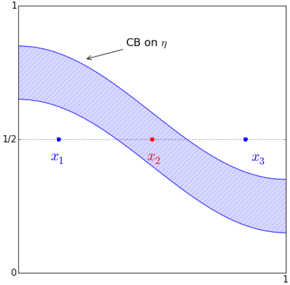
$$\text{For } \beta = 1: \quad \inf_{\hat{f}_n} \sup_{\eta} \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

- Unrestricted $\mathbb{P}_X$: different minimax rate

$$\text{Active}: \Theta\left(n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}\right) \text{ vs. Passive}: \Theta\left(n^{-\frac{\alpha(\beta+1)}{2\alpha+d+\alpha\beta}}\right)$$

Our results: algorithmic contribution

Naive strategy: suppose we have a Confidence Band on η



Request new label at x_2 but not at x_1, x_3

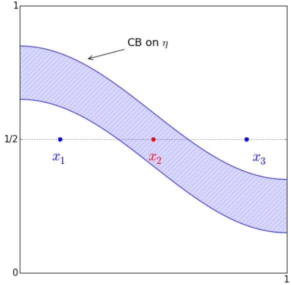
Optimal CBs require strong conditions on η (e.g. self-similarity)

New generic adaptation strategy for nested classes $\{\Sigma(\alpha)\}_{\alpha>0}$

Aggregate $\hat{Y}$ estimates from non-adaptive subroutines (over $\alpha \nearrow$).

Our results: algorithmic contribution

Naive strategy: suppose we have a Confidence Band on η



Request new label at x_2 but not at x_1, x_3

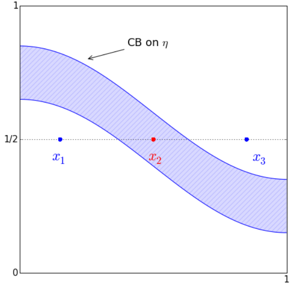
Optimal CBs require strong conditions on η (e.g. self-similarity)

New generic adaptation strategy for nested classes $\{\Sigma(\alpha)\}_{\alpha>0}$

Aggregate $\hat{Y}$ estimates from non-adaptive subroutines (over $\alpha \nearrow$).

Our results: algorithmic contribution

Naive strategy: suppose we have a Confidence Band on η



Request new label at x_2 but not at x_1, x_3

Optimal CBs require strong conditions on η (e.g. self-similarity)

New generic adaptation strategy for nested classes $\{\Sigma(\alpha)\}_{\alpha>0}$

Aggregate $\hat{Y}$ estimates from non-adaptive subroutines (over $\alpha \nearrow$).

Outline

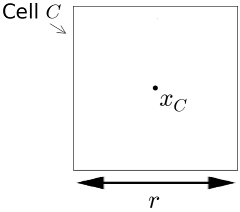
- **Upper-bounds**
  - **Non-adaptive Subroutine**
  - Adaptive Procedure
- Lower-bounds

Non-adaptive Subroutine

Suppose we know η is α-smooth $(\alpha \le 1)$

- Query t labels at x_C and estimate $\eta(x_C)$:

$$\text{w.h.p.} \quad |\widehat{\eta}(x_C) - \eta(x_C)| \lesssim \sqrt{\frac{1}{t}}$$



- We know η changes on C by at most r^α

$$\implies \forall x \in C, \quad |\widehat{\eta}(x_C) - \eta(x)| \lesssim \sqrt{\tfrac{1}{t}} + r^\alpha$$

$\therefore$ Let $t \approx r^{-2\alpha}$, we can safely label C if

$$\boxed{|\widehat{\eta}(x_C) - 1/2| \gtrsim 2r^\alpha}$$

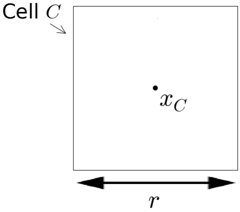
Otherwise partition C and repeat over smaller regions.

Non-adaptive Subroutine

Suppose we know η is α-smooth $(\alpha \leq 1)$

- Query t labels at x_C and estimate $\eta(x_C)$:

$$\text{w.h.p.} \quad |\widehat{\eta}(x_C) - \eta(x_C)| \lesssim \sqrt{\frac{1}{t}}$$



- We know η changes on C by at most r^α

$$\implies \forall x \in C, \quad |\widehat{\eta}(x_C) - \eta(x)| \lesssim \sqrt{\tfrac{1}{t}} + r^\alpha$$

$\therefore$ Let $t \approx r^{-2\alpha}$, we can safely label C if

$$|\widehat{\eta}(x_C) - 1/2| \gtrsim 2r^\alpha$$

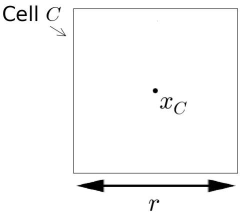
Otherwise partition C and repeat over smaller regions.

Non-adaptive Subroutine

Suppose we know η is α-smooth $(\alpha \leq 1)$

- Query t labels at x_C and estimate $\eta(x_C)$:

$$\text{w.h.p.} \quad |\widehat{\eta}(x_C) - \eta(x_C)| \lesssim \sqrt{\frac{1}{t}}$$



- We know η changes on C by at most r^α

$$\implies \forall x \in C, \quad |\widehat{\eta}(x_C) - \eta(x)| \lesssim \sqrt{\tfrac{1}{t}} + r^\alpha$$

$\therefore$ Let $t \approx r^{-2\alpha}$, we can safely label C if

$$\boxed{|\widehat{\eta}(x_C) - 1/2| \gtrsim 2r^\alpha}$$

Otherwise partition C and repeat over smaller regions.

Non-adaptive Subroutine

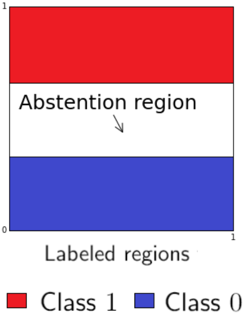
Suppose we know η is α-smooth ($\alpha \leq 1$)

Implement previous intuition over **hierarchical partition** of $[0,1]^d$.

Final output given budget n:

- Correctly labeled subset of $[0,1]^d$
- Abstention region contained in $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}\}$.

$\Delta_{\alpha,\beta} \doteq \Delta_{\alpha,\beta}(n)$ is "optimal" under different $\mathbb{P}_X$ regimes.



Case $\alpha > 1$:
Same intuition, but higher order interpolation (for $\hat{\eta}$) on cells C

Non-adaptive Subroutine

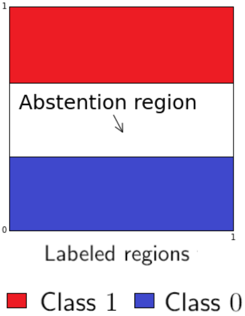
Suppose we know η is α-smooth ($\alpha \leq 1$)

Implement previous intuition over **hierarchical partition** of $[0,1]^d$.

Final output given budget n:

- Correctly labeled subset of $[0,1]^d$
- Abstention region contained in $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}\}$.

  $\Delta_{\alpha,\beta} \doteq \Delta_{\alpha,\beta}(n)$ is "optimal" under different $\mathbb{P}_X$ regimes.



Case $\alpha > 1$:
Same intuition, but higher order interpolation (for $\hat{\eta}$) on cells C

Non-adaptive Subroutine

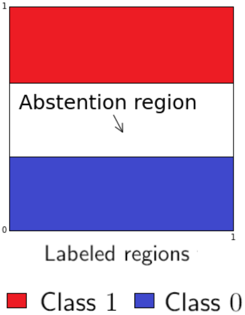
Suppose we know η is α-smooth $(\alpha \leq 1)$

Implement previous intuition over **hierarchical partition** of $[0,1]^d$.

Final output given budget n:

- Correctly labeled subset of $[0,1]^d$
- Abstention region contained in $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}\}$.

$\Delta_{\alpha,\beta} \doteq \Delta_{\alpha,\beta}(n)$ is "optimal" under different $\mathbb{P}_X$ regimes.



Case $\alpha > 1$:
Same intuition, but higher order interpolation (for $\hat{\eta}$) on cells C

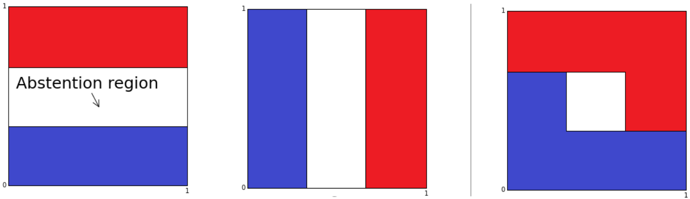
Outline

- **Upper-bounds**
  - Non-adaptive Subroutine
  - **Adaptive Procedure**
- Lower-bounds

Adaptive Procedure (α unknown)

Key idea: η is α'-Hölder for any $\alpha' \leq \alpha$
$\implies$ Subroutine(α') returns correct labels (red or blue)

Procedure:
Aggregate labelings of Subroutine(α') for $\alpha' = \alpha_1 < \alpha_2 < \ldots$



Labeled regions for α_1 | Labeled regions for α_2 | Aggregated labels

Correctness: at $\alpha_i = \alpha$ labeling has optimal error
At $\alpha_i > \alpha$, we never overwrite previous labels (error remains small)

Implementation: $\alpha_i \in \left[\frac{1}{\log n} : \frac{1}{\log n} : \log n\right]$, use budget $\frac{n}{\log^2 n}$ $\forall \alpha_i$

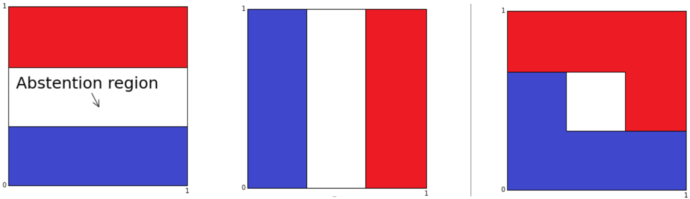
Adaptive Procedure (α unknown)

Key idea: η is α'-Hölder for any $\alpha' \leq \alpha$

$\implies$ Subroutine(α') returns correct labels (red or blue)

Procedure:

Aggregate labelings of Subroutine(α') for $\alpha' = \alpha_1 < \alpha_2 < \ldots$



Labeled regions for α_1 — Labeled regions for α_2 — Aggregated labels

Correctness: at $\alpha_i = \alpha$ labeling has optimal error

At $\alpha_i > \alpha$, we never overwrite previous labels (error remains small)

Implementation: $\alpha_i \in \left[\frac{1}{\log n} : \frac{1}{\log n} : \log n\right]$, use budget $\frac{n}{\log^2 n}$ $\forall \alpha_i$

Adaptive Procedure (α unknown)

Without self-similarity assumptions adaptive $\widehat{f}_n$ satisfies:

Theorem: unrestricted $\mathbb{P}_X$

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem: $\mathbb{P}_X$ uniform

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-(\alpha\wedge 1)\beta}}$$

which are all tight rates.

Adaptive Procedure (α unknown)

Without self-similarity assumptions adaptive $\widehat{f}_n$ satisfies:

Theorem: unrestricted $\mathbb{P}_X$

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem: $\mathbb{P}_X$ uniform

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-(\alpha\wedge 1)\beta}}$$

which are all tight rates.

Adaptive Procedure (α unknown)

Without self-similarity assumptions adaptive $\widehat{f}_n$ satisfies:

Theorem: unrestricted $\mathbb{P}_X$

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem: $\mathbb{P}_X$ uniform

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-(\alpha\wedge 1)\beta}}$$

which are all tight rates.

Outline

- Upper-bounds
  - Non-adaptive Subroutine
  - Adaptive Procedure
- **Lower-bounds**

Lower-bounds

Theorem (unrestricted $\mathbb{P}_X$)

For any active learner $\hat{f}_n$ we have:

$$\sup_\eta \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq Cn^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem ($\mathbb{P}_X$ uniform and $\alpha > 1$, $\beta = 1$)

For any active learner $\hat{f}_n$ we have:

$$\sup_\eta \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq Cn^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

This confirms a transition in the rate (at least for $\beta = 1$).

Lower-bounds

Theorem (unrestricted $\mathbb{P}_X$)

For any active learner $\hat{f}_n$ we have:

$$\sup_{\eta} \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem ($\mathbb{P}_X$ uniform and $\alpha > 1$, $\beta = 1$)

For any active learner $\hat{f}_n$ we have:

$$\sup_{\eta} \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

This confirms a transition in the rate (at least for $\beta = 1$).

Lower-bounds

Theorem (unrestricted $\mathbb{P}_X$)

For any active learner $\hat{f}_n$ we have:

$$\sup_\eta \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d}}$$

Theorem ($\mathbb{P}_X$ uniform and $\alpha > 1$, $\beta = 1$)

For any active learner $\hat{f}_n$ we have:

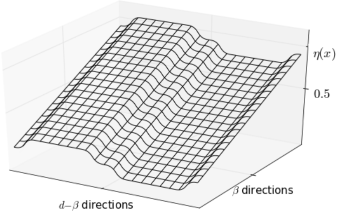
$$\sup_\eta \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

This confirms a transition in the rate (at least for $\beta = 1$).

Lower-bound construction for $\mathbb{P}_X$ uniform, $\alpha > 1$, $\beta = 1$

Remember difference in rates:
$\alpha \leq 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}}$
$\alpha > 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$

Hard case for $\alpha > 1$:
η changes linearly in β directions,
but oscillates in $d - \beta$ directions



...$d - \beta$ now acts as the effective degrees of freedom

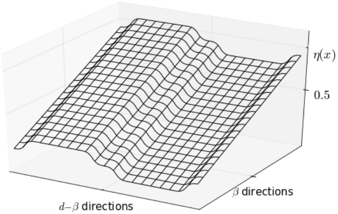
Lower-bound construction for $\mathbb{P}_X$ uniform, $\alpha > 1$, $\beta = 1$

Remember difference in rates:

$\alpha \leq 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}}$

$\alpha > 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$

Hard case for $\alpha > 1$:
η changes linearly in β directions,
but oscillates in $d - \beta$ directions



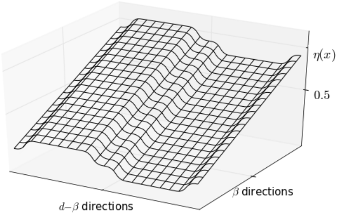
...$d - \beta$ now acts as the effective degrees of freedom

Lower-bound construction for $\mathbb{P}_X$ uniform, $\alpha > 1$, $\beta = 1$

Remember difference in rates:

$\alpha \leq 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\alpha\beta}}$

$\alpha > 1 : n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$



Hard case for $\alpha > 1$:
η changes linearly in β directions,
but oscillates in $d - \beta$ directions

...$d - \beta$ now acts as the effective degrees of freedom

Summary

- We recover rates in A-L under more natural assumptions
- Confirmed a conjectured transition at $\alpha > 1$
- Established new minimax rates for unrestricted $\mathbb{P}_X$
- Introduced a generic adaptation framework for nested classes

Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Summary

- We recover rates in A-L under more natural assumptions
- Confirmed a conjectured transition at $\alpha > 1$
- Established new minimax rates for unrestricted $\mathbb{P}_X$
- Introduced a generic adaptation framework for nested classes

Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Summary

- We recover rates in A-L under more natural assumptions
- Confirmed a conjectured transition at $\alpha > 1$
- Established new minimax rates for unrestricted $\mathbb{P}_X$
- Introduced a generic adaptation framework for nested classes

Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Summary

- We recover rates in A-L under more natural assumptions
- Confirmed a conjectured transition at $\alpha > 1$
- Established new minimax rates for unrestricted $\mathbb{P}_X$
- Introduced a generic adaptation framework for nested classes

Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Summary

- We recover rates in A-L under more natural assumptions
- Confirmed a conjectured transition at $\alpha > 1$
- Established new minimax rates for unrestricted $\mathbb{P}_X$
- Introduced a generic adaptation framework for nested classes

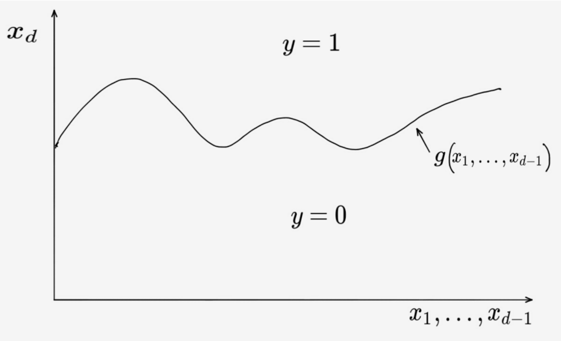
Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Outline:

We consider various regularity conditions on $\eta = \mathbb{E}\left[Y|X\right]$:

- η nearly aligns with clusters in X
  with R. Urner and S. Ben David, 2015
- blue η is a smooth function
  with A. Locatelli and A. Carpentier, 2017
- η defines a smooth decision-boundary
  with A. Locatelli and A. Carpentier, soon on Arxiv

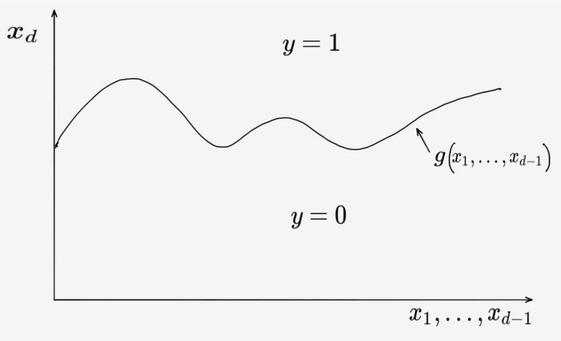
η defines a smooth decision-boundary



- $\mathcal{D} \equiv \{x : \eta(x) = 1/2\}$ is given by α-Hölder function g.
- **Noise condition:** $|\eta(x) - 1/2| \approx \text{dist}(x, \mathcal{D})^{\kappa-1}$, $\kappa > 1$.

Problem is easier as $\kappa \to 1, \alpha \to \infty$.

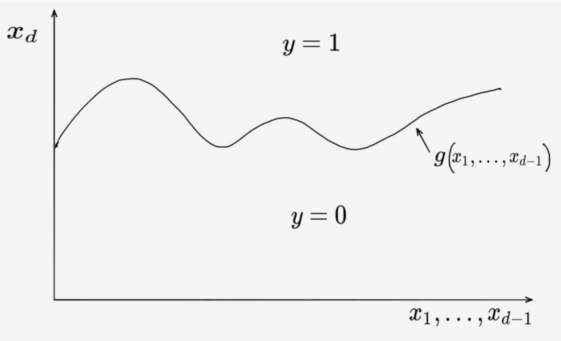
η defines a smooth decision-boundary



- $\mathcal{D} \equiv \{x : \eta(x) = 1/2\}$ is given by α-Hölder function g.
- **Noise condition:** $|\eta(x) - 1/2| \approx \text{dist}(x, \mathcal{D})^{\kappa-1}$, $\kappa > 1$.

Problem is easier as $\kappa \to 1, \alpha \to \infty$.

η *defines a smooth decision-boundary*



- $\mathcal{D} \equiv \{x : \eta(x) = 1/2\}$ is given by α-Hölder function g.
- **Noise condition:** $|\eta(x) - 1/2| \approx \text{dist}(x, \mathcal{D})^{\kappa - 1}$, $\kappa > 1$.

Problem is easier as $\kappa \to 1, \alpha \to \infty$.

Previous work [Castro, Nowak 07], $P_X \equiv \mathcal{U}[0,1]^d$

If we know α, κ, then:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha\kappa}{2\alpha(\kappa-1)+d-1}} \quad \text{(rate is tight)}$$

Passive rate: Replace $\kappa - 1$ with $\kappa - 1/2$.

Can these gains be achieved by an adaptive procedure?

Previous work [Castro, Nowak 07], $P_X \equiv \mathcal{U}[0,1]^d$

If we know α, κ*, then:*

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha\kappa}{2\alpha(\kappa-1)+d-1}} \quad \text{(rate is tight)}$$

Passive rate: Replace $\kappa - 1$ with $\kappa - 1/2$.

Can these gains be achieved by an adaptive procedure?

Previous work [Castro, Nowak 07], $P_X \equiv \mathcal{U}[0,1]^d$

If we know α, κ, then:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha\kappa}{2\alpha(\kappa-1)+d-1}} \quad \text{(rate is tight)}$$

Passive rate: Replace $\kappa - 1$ with $\kappa - 1/2$.

Can these gains be achieved by an adaptive procedure?

Previous work [Castro, Nowak 07], $P_X \equiv \mathcal{U}[0,1]^d$

If we know α, κ, then:

$$R(\hat{f}_n) - R(f^*) \lesssim n^{-\frac{\alpha\kappa}{2\alpha(\kappa-1)+d-1}} \quad \text{(rate is tight)}$$

Passive rate: Replace $\kappa - 1$ with $\kappa - 1/2$.

Can these gains be achieved by an adaptive procedure?

Existing adaptive results:

Dimension $d = 1$, $\mathcal{D} \equiv$ threshold on the line

Binary search strategies are adaptive to κ ... (fixed $\alpha = \infty$)
[Hanneke, 09], [Ramdas, Singh 13], [Yan, Chaudhuri, Javidi, 16]

Use any of these (blackbox) to get a fully adaptive strategy in $\mathbb{R}^d$!

Existing adaptive results:

Dimension $d = 1$, $\mathcal{D} \equiv$ threshold on the line

Binary search strategies are adaptive to κ ... (fixed $\alpha = \infty$)

[Hanneke, 09], [Ramdas, Singh 13], [Yan, Chaudhuri, Javidi, 16]

Use any of these (blackbox) to get a fully adaptive strategy in $\mathbb{R}^d$!

Intuition:
If $\mathcal{D}$ is α-smooth, then it's α'-smooth for $\alpha' \leq \alpha$!

So use the same strategy as before:

Aggregate estimates from non-adaptive subroutine for $\alpha \nearrow$

Main difficulty: such subroutine must adapt to κ in $\mathbb{R}^d$ ...

Intuition:
If $\mathcal{D}$ is α-smooth, then it's α'-smooth for $\alpha' \leq \alpha$!

So use the same strategy as before:

Aggregate estimates from non-adaptive subroutine for $\alpha \nearrow$

Main difficulty: such subroutine must adapt to κ in $\mathbb{R}^d$ ...

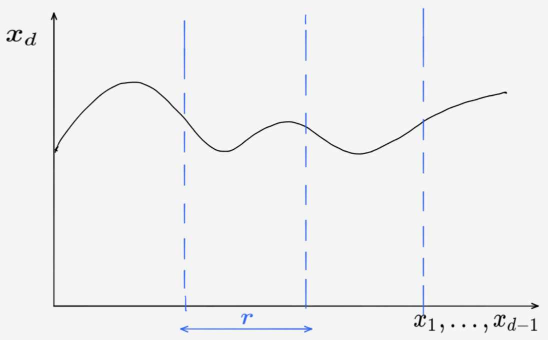
Intuition:
If $\mathcal{D}$ is α-smooth, then it's α'-smooth for $\alpha' \leq \alpha$!

So use the same strategy as before:

Aggregate estimates from non-adaptive subroutine for $\alpha \nearrow$

Main difficulty: such subroutine must adapt to κ in $\mathbb{R}^d$ ...

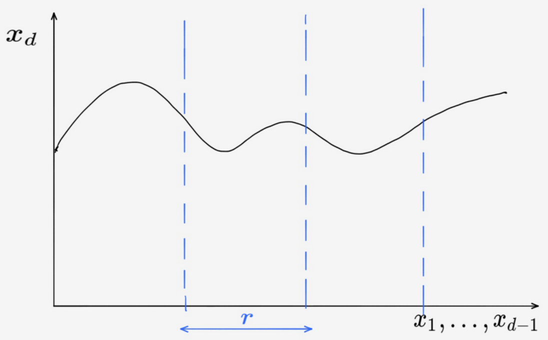
SubRoutine: suppose α were known



Partition $[0,1]^{d-1}$ into cells of side-length r.

...

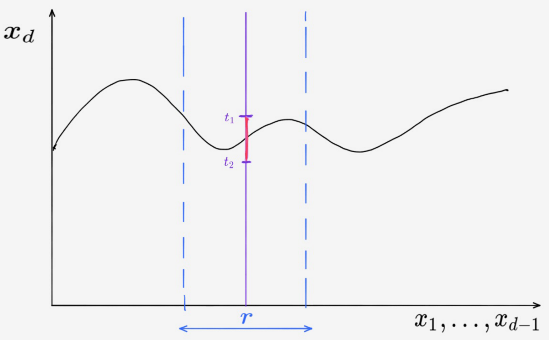
SubRoutine: suppose α were known



Partition $[0,1]^{d-1}$ into cells of side-length r.

$\ldots$

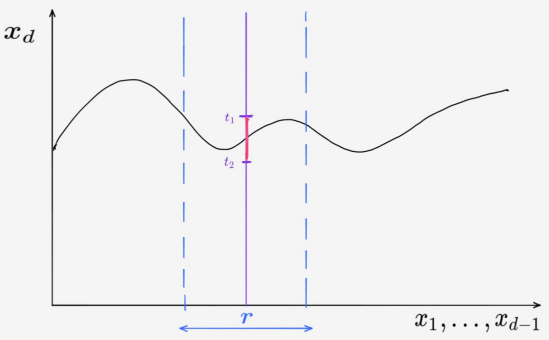
Subroutine: suppose α were known



Line search in each cell returns $[t_1, t_2]$ intersecting $\mathcal{D}$.

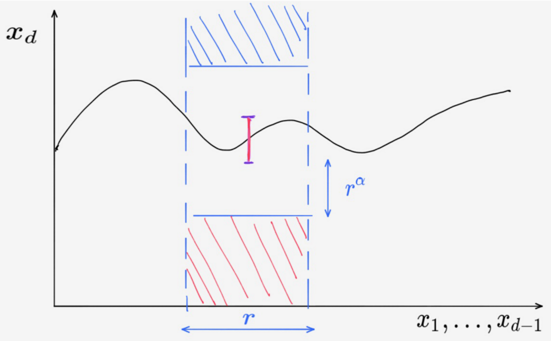
$|t_2 - t_1|$ is optimal in terms of unknown κ ...

Subroutine: suppose α were known



Line search in each cell returns $[t_1, t_2]$ intersecting $\mathcal{D}$.
$|t_2 - t_1|$ is optimal in terms of unknown κ ...

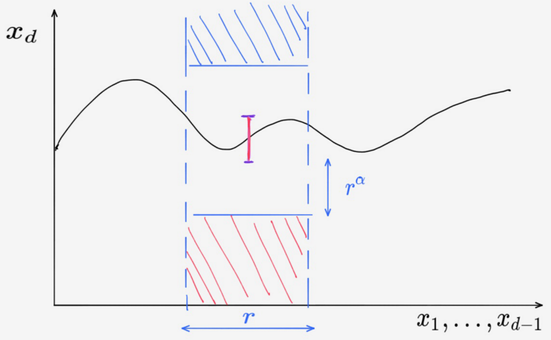
Subroutine: suppose α were known



$\alpha \leq 1$: We know $\mathcal{D}$ is at most r^α away through the cell

$\alpha > 1$: use more careful (higher-order) extrapolation.

Subroutine: suppose α were known

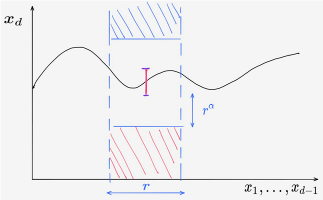


$\alpha \leq 1$: We know $\mathcal{D}$ is at most r^α away through the cell
$\alpha > 1$: use more careful (higher-order) extrapolation.

Subroutine: suppose α were known.

Aggregate over $r \in [\frac{1}{2}, \frac{1}{4}, \ldots, 1/n]$:

Final labeling is optimal w.r.t. κ, α



Active learning procedure: (adapting to α)

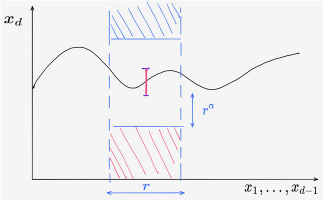
Call subroutine for $\alpha_i \in \left[\frac{1}{\log n} : \frac{1}{\log n} : \log n\right]$, use budget $\frac{n}{\log^2 n}$ $\forall \alpha_i$.

We then get the first fully adaptive and optimal A-L for the setting!

Subroutine: suppose α were known.

Aggregate over $r \in [\frac{1}{2}, \frac{1}{4}, \ldots, 1/n]$:

Final labeling is optimal w.r.t. κ, α



Active learning procedure: (adapting to α)

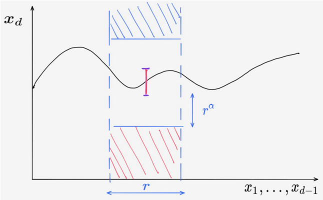
Call subroutine for $\alpha_i \in \left[\frac{1}{\log n} : \frac{1}{\log n} : \log n\right]$, use budget $\frac{n}{\log^2 n}$ $\forall \alpha_i$.

We then get the first fully adaptive and optimal A-L for the setting!

Subroutine: suppose α were known.

Aggregate over $r \in [\frac{1}{2}, \frac{1}{4}, \ldots, 1/n]$:

Final labeling is optimal w.r.t. κ, α



Active learning procedure: (adapting to α)

Call subroutine for $\alpha_i \in \left[\frac{1}{\log n} : \frac{1}{\log n} : \log n\right]$, use budget $\frac{n}{\log^2 n}$ $\forall \alpha_i$.

We then get the first fully adaptive and optimal A-L for the setting!

In summary:

Further gains in A-L emerge as we parametrize from easy to hard.

There is much left to understand ...

τ

Thanks!

In summary:

Further gains in A-L emerge as we parametrize from easy to hard.
There is much left to understand ...

τ

Thanks!

In summary:

Further gains in A-L emerge as we parametrize from easy to hard.
There is much left to understand ...
τ

Thanks!