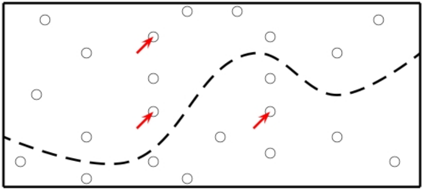
Adaptive Strategies for Nonparametric Active Learning

Andrea Locatelli

(Uni Magdeburg)

Based on works with Alexandra Carpentier and Samory Kpotufe

Active Classification



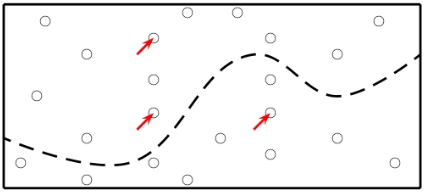
Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

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



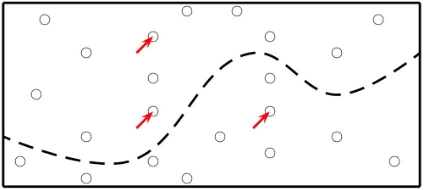
Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

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



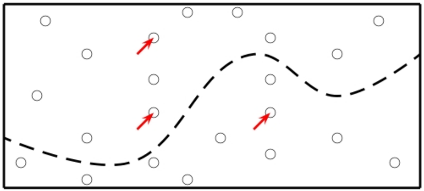
Pb: Classification $X \to Y \in \{0, 1\}$ when **labels are expensive**.
Goal: Return a good classifier using **few label queries.**

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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

But $R(f^*)$ is often > 0 (imperfect world):
noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

noisy images or speech, adversarial spam, variable 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]

A-L rates $\equiv \sqrt{R(f^*)/n} + e^{-\sqrt{n}}$, vs P-L rates $\equiv \sqrt{R(f^*)/n} + 1/n$

$R(f^*) > 0$: both rates are $\equiv 1/\sqrt{n}$ (no significant gain).

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

Are there no gains in these practical settings?

We want to understand which gains are possible over passive learning under general conditions, and for reasonable procedures.

General Conditions:

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. Tsybakov, Massart)

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

General Conditions:

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. Tsybakov, Massart)

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

General Conditions:

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. Tsybakov, Massart)

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

General Conditions:

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. Tsybakov, Massart)

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

General Conditions:

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. Tsybakov, Massart)

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

General Conditions:

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. Tsybakov, Massart)

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

Initial insights ... different regularity conditions

[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:

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

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

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on Disagreement Coefficient!

However:
Needs full knowledge of boundary regularity and noise decay.

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

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on Disagreement Coefficient!

However:
Needs full knowledge of boundary regularity and noise decay.

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

Show considerable gains over passive learning even with label noise!
Implementable, no conditions on Disagreement Coefficient!

However:

Needs full knowledge of boundary regularity and noise decay.

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise! **Implementable, no conditions on Disagreement Coefficient, Adaptive**!

However:

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

[Minsker, 2012] (η is smooth):

Show considerable gains over passive learning even with label noise! **Implementable, no conditions on Disagreement Coefficient, Adaptive**!

However:

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

Can reasonable A-L procedures **(implementable + adaptive)** attain considerable gains over P-L for **general distributions**?

Some of our recent results:

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

- η is a smooth function
  with A. Carpentier and S.Kpotufe, COLT 2017
- η defines a smooth decision-boundary
  with S.Kpotufe and A. Carpentier, ALT 2018

Outline:

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

- η is a smooth function
  with A. Carpentier and S.Kpotufe, COLT 2017
- η defines a smooth decision-boundary
  with S.Kpotufe and A. Carpentier, ALT 2018

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)

  Example: $\alpha = 1 \implies \eta$ is Lipschitz.

- Tsybakov noise condition: $\exists c, \beta \geq 0$ such that $\forall \tau > 0$:

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

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)
  Example: $\alpha = 1 \implies \eta$ is Lipschitz.

- Tsybakov noise condition: $\exists c, \beta \geq 0$ such that $\forall \tau > 0$:

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

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)

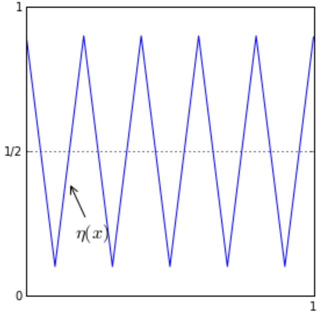
  Example: $\alpha = 1 \implies \eta$ is Lipschitz.

- Tsybakov noise condition: $\exists c, \beta \geq 0$ such that $\forall \tau > 0$:

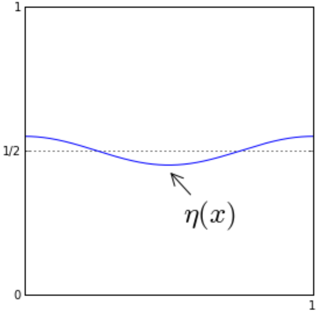
$$\mathbb{P}_X \left(x : \left| \eta(x) - \frac{1}{2} \right| \leq \tau \right) \leq c\tau^{\beta},$$

...

α, β capture continuum between easy and hard problems



Small α



Small β

...

α, β capture continuum between easy and hard problems

[Audibert-Tsybakov 07]

Passive rates : $n^{-(\beta+1)/\left(2+\frac{d}{\alpha}\right)}$

The above implies:

- **Slow rates of** $\Omega(n^{-1/d})$ for small α, β.
- **Fast rates of** $o(1/n)$: for large α, β.

...

α, β capture continuum between easy and hard problems

[Audibert-Tsybakov 07]

Passive rates : $n^{-(\beta+1)/\left(2+\frac{d}{\alpha}\right)}$

The above implies:

- **Slow rates of** $\Omega(n^{-1/d})$ for small α, β.
- **Fast rates of** $o(1/n)$: for large α, β.

We'll see that: interaction between α, β and d control A-L rates

...

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

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

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

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

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

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

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

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

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

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

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

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

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

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

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

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

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

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

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

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

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

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

Theorem: $\alpha \le 1$, $\alpha\beta < d$

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 $\alpha > 1$ Minsker conjectures a transition:

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

Open: Unrestricted $\mathbb{P}_X$? General η? $\alpha\beta = d$? $\alpha > 1$?

We'll present both new **statistical** and **algorithmic** results:

Statistical contributions

Significantly milder conditions, new rate regimes:

- Recover all rates without self-similarity conditions on η.
- $\mathbb{P}_X$ uniform (new transitions):
  - No (exponential) dependence on d when $\min\{\alpha, 1\}\beta = 1$.
  - Verify rate transition for $\alpha > 1$:

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

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

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

Statistical contributions

Significantly milder conditions, new rate regimes:

- Recover all rates without self-similarity conditions on η.
- $\mathbb{P}_X$ uniform (new transitions):
  - No (exponential) dependence on d when $\min\{\alpha, 1\}\beta = 1$.
  - Verify rate transition for $\alpha > 1$:

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

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

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

Statistical contributions

Significantly milder conditions, new rate regimes:

- Recover all rates without self-similarity conditions on η.
- $\mathbb{P}_X$ uniform (new transitions):
  - No (exponential) dependence on d when $\min\{\alpha, 1\}\beta = 1$.
  - Verify rate transition for $\alpha > 1$:

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

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

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

Statistical contributions

Significantly milder conditions, new rate regimes:

- Recover all rates without self-similarity conditions on η.
- $\mathbb{P}_X$ uniform (new transitions):
  - No (exponential) dependence on d when $\min\{\alpha, 1\}\beta = 1$.
  - Verify rate transition for $\alpha > 1$:

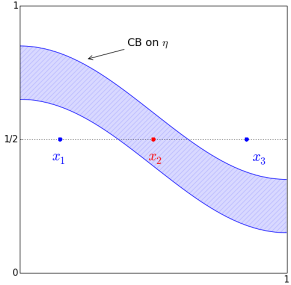
$$\text{For } \beta = 1 : \quad \inf_{\hat{f}_n} \sup_{\eta} \mathbb{E}[R(\hat{f}_n)] - R(f^*) \gtrsim n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

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

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

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

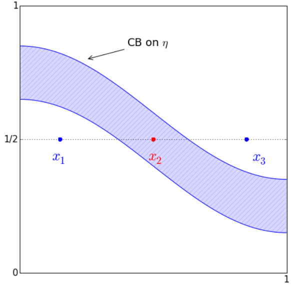
Optimal CBs require strong conditions on η (e.g. self-similarity)

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

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

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

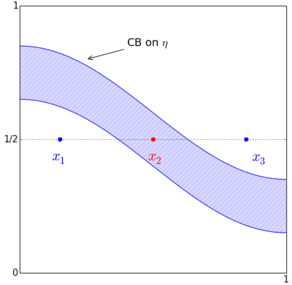
Optimal CBs require strong conditions on η (e.g. self-similarity)

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

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

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

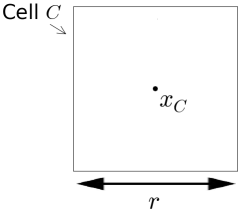
- **Upper-bounds**
  - **Non-adaptive Subroutine**
  - Adaptive Procedure
- Lower-bounds

Non-adaptive Subroutine

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

- We know η changes on C by at most r^α
- Query t labels at x_C and estimate $\eta(x_C)$:

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



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

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

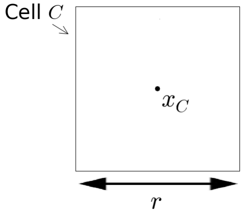
$$\boxed{|\widehat{\eta}(x_C) - 1/2| \gtrsim 2r^\alpha}$$

Otherwise partition C and repeat over smaller regions.

Non-adaptive Subroutine

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

- We know η changes on C by at most r^α
- Query t labels at x_C and estimate $\eta(x_C)$:



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

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

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

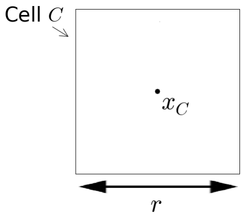
$$|\widehat{\eta}(x_C) - 1/2| \gtrsim 2r^\alpha$$

Otherwise partition C and repeat over smaller regions.

Non-adaptive Subroutine

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

- We know η changes on C by at most r^α
- Query t labels at x_C and estimate $\eta(x_C)$:



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

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

$\therefore$ 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

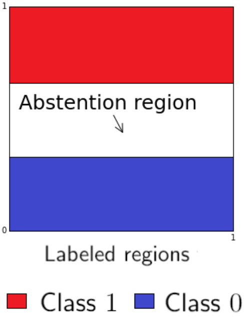
Suppose we know η is α-smooth $(\alpha \leq 1)$

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 $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}(n)\}$.

$\Delta_{\alpha,\beta}(n)$ 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

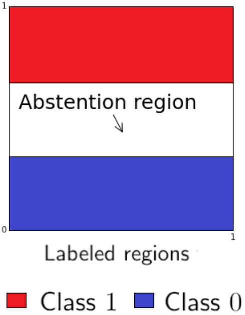
Suppose we know η is α-smooth $(\alpha \leq 1)$

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 $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}(n)\}$.

  $\Delta_{\alpha,\beta}(n)$ 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

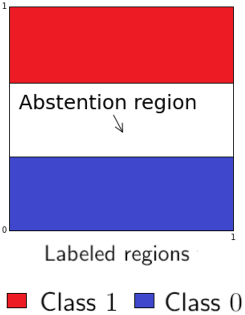
Suppose we know η is α-smooth $(\alpha \leq 1)$

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 $\{x : |\eta(x) - 1/2| \leq \Delta_{\alpha,\beta}(n)\}$.

$\Delta_{\alpha,\beta}(n)$ 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)

Difficulty: Collected labels depend on parameters of A-L algorithm

First idea: Split budget and cross-validate over values of α ...
Cost: (optimal rate) + $1/\sqrt{n}$

So cannot get fast rates ...

Adaptive Procedure (α unknown)

Difficulty: Collected labels depend on parameters of A-L algorithm

First idea: Split budget and cross-validate over values of α ...
Cost: (optimal rate) + $1/\sqrt{n}$

So cannot get fast rates ...

Adaptive Procedure (α unknown)

Difficulty: Collected labels depend on parameters of A-L algorithm

First idea: Split budget and cross-validate over values of α ...
Cost: (optimal rate) + $1/\sqrt{n}$

So cannot get fast rates ...

Adaptive Procedure (α unknown)

Difficulty: Collected labels depend on parameters of A-L algorithm

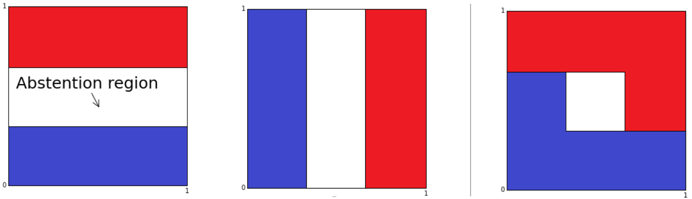
First idea: Split budget and cross-validate over values of α ...
Cost: (optimal rate) $+ 1/\sqrt{n}$

So cannot get fast rates ...

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

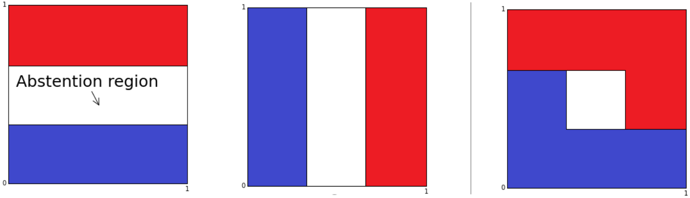
Correctness: at $\alpha_i = \alpha$ labeling has optimal error
At $\alpha_i > \alpha$, we never overwrite previous labels (error remains small)

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 $\alpha' = \alpha_1 < \alpha_2 < \dots$



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.

Outline

- η is a smooth function
  with A. Carpentier and S.Kpotufe, COLT 2017
  - Upper-bounds
    - Non-adaptive Subroutine
    - Adaptive Procedure
  - **Lower-bounds**
- η defines a smooth decision-boundary
  with S.Kpotufe and A. Carpentier, ALT 2018

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:

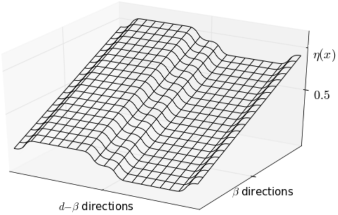
$$\sup_{\eta} \mathbb{E}[R(\hat{f}_n)] - R(f^*) \geq C n^{-\frac{\alpha(\beta+1)}{2\alpha+d-\beta}}$$

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

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



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

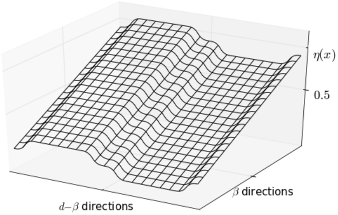
Lower-bound construction for $\mathbb{P}_X$ uniform, $\alpha > 1$, $\beta = 1$

Remember difference in rates:

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

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

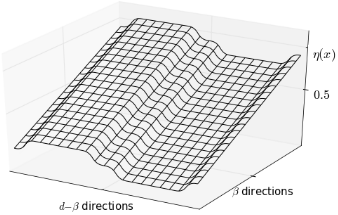
Hard case for $\alpha > 1$, $\beta = 1$:
η changes linearly in 1 direction,
but oscillates in $d-1$ directions



...$d - \beta$ 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 $\alpha > 1$, $\beta = 1$:
η changes linearly in 1 direction,
but oscillates in $d - 1$ directions

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

Summary

- We recover rates in A-L under more natural assumptions
- Different transitions: $\alpha > 1$, $(\alpha \wedge 1)\beta = d$, 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
- Different transitions: $\alpha > 1$, $(\alpha \wedge 1)\beta = d$, 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
- Different transitions: $\alpha > 1$, $(\alpha \wedge 1)\beta = d$, 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
- Different transitions: $\alpha > 1$, $(\alpha \wedge 1)\beta = d$, unrestricted $\mathbb{P}_X$.
- Introduced a generic adaptation framework for nested classes.

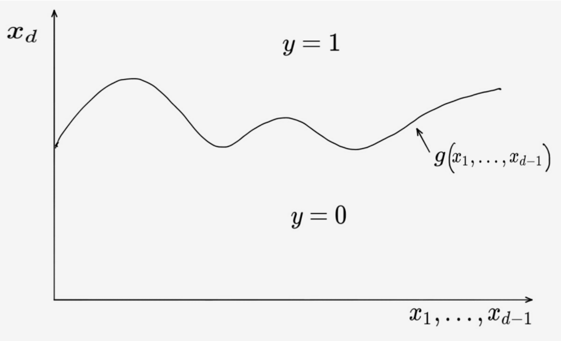
Extension: our framework yields the first adaptive procedure in the smooth boundary setting of Castro and Nowak (2008)

Our recent result:

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

- η is a smooth function
  with A. Carpentier and S. Kpotufe, COLT 2017
- η defines a smooth decision-boundary
  with S.Kpotufe and A. Carpentier, ALT 2018

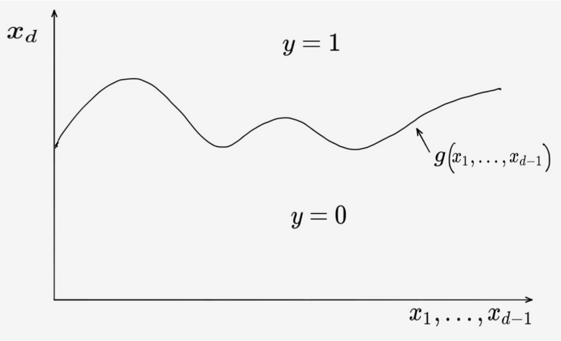
η *defines a smooth decision-boundary*



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

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

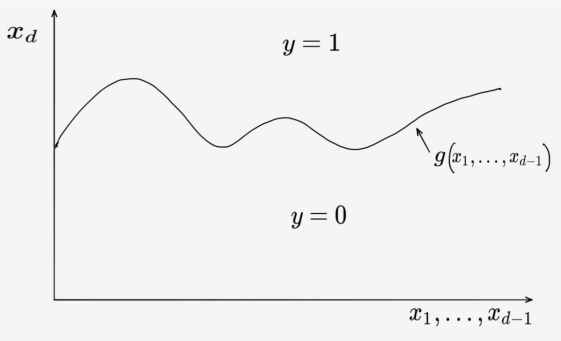
η *defines a smooth decision-boundary*



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

Problem gets 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:** $|\eta(x) - 1/2| \approx \text{dist}(x, \mathcal{D})^{\kappa-1}$, $\kappa \geq 1$.

Problem gets 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]

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:

- Subroutine must **adapt** to κ in $\mathbb{R}^d$ ...
- Subroutine must **estimate** boundary optimally...
- Use α to **abstain** from labeling when unsure...

Our subroutine builds on a known reduction to line search

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:

- Subroutine must **adapt** to κ in $\mathbb{R}^d$ ...
- Subroutine must **estimate** boundary optimally...
- Use α to **abstain** from labeling when unsure...

Our subroutine builds on a known reduction to line search

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:

- Subroutine must **adapt** to κ in $\mathbb{R}^d$ ...
- Subroutine must **estimate** boundary optimally...
- Use α to **abstain** from labeling when unsure...

Our subroutine builds on a known reduction to line search

Intuition:

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

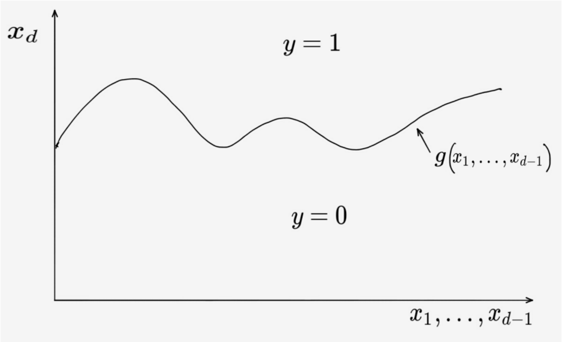
So use the same strategy as before:

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

Main difficulty:

- Subroutine must **adapt** to κ in $\mathbb{R}^d$ ...
- Subroutine must **estimate** boundary optimally...
- Use α to **abstain** from labeling when unsure...

Our subroutine builds on a known reduction to line search



We 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.

Next directions:

- Better aggregation?
- Draw links with Contextual Bandits, Nonlinear Optimization.

Thanks!

In summary:

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

Next directions:

- Better aggregation?
- Draw links with Contextual Bandits, Nonlinear Optimization.

Thanks!