Adaptive inference and its relations to sequential decision making

Alexandra Carpentier[1]

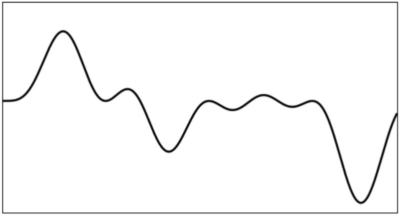
OvGU Magdeburg

Based on joint works with **Olga Klopp, Samory Kpotufe, Andréa Locatelli, Matthias Löffler, Richard Nickl**

Criteo, Oct. 2nd, 2019

[1]Partly funded by the DFG EN CA1488, the CRC 1294, the GK 2297, the GK 2433

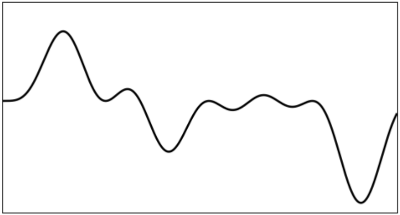
Non-Convex Optimization



Problem

Finding/Exploiting the maximum $M(f)$ of an unknown function f.

Non-Convex Optimization



Problem

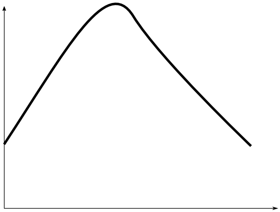
Finding/Exploiting the maximum $M(f)$ of an unknown function f.

Question

Can we design algorithms that adapt to the difficulty of the problem?

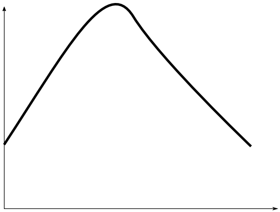
Non-Convex Optimization

Depending on the difficulty of the problem, we would hope to get different performances :



Non-Convex Optimization

Depending on the difficulty of the problem, we would hope to get different performances :



Question

Can we adapt to the hyperparameters?

Scope of this talk

Talk :

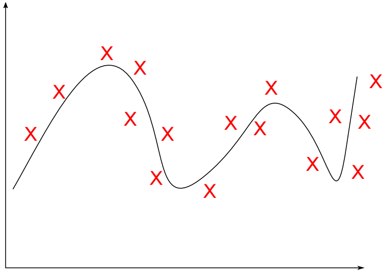
- Presentation of adaptive inference in statistics.
- Adaptivity in continuously armed bandits.

ADAPTIVE INFERENCE

Adaptive inference for non-parametric regression

Problem :

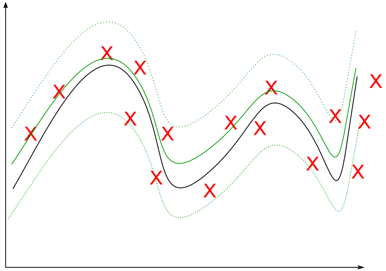
Non-parametric regression



Inference (estimation + uncertainty quantification) of the function?

Adaptive inference for non-parametric regression

Problem :
Non-parametric regression

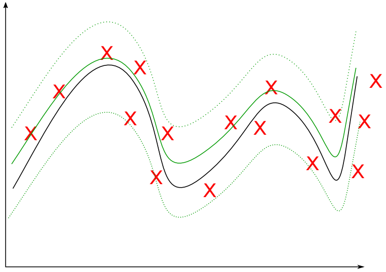


Inference (estimation + uncertainty quantification) of the function?

Adaptive inference for non-parametric regression

Problem :

Non-parametric regression



Inference (estimation + uncertainty quantification) of the function?

The Model

f: function on $[0,1]^d$.

n observed data samples $(X_i, Y_i)_{i \le n}$:

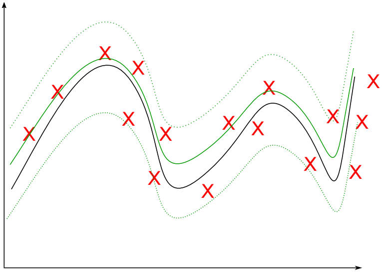
$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{[0,1]^d}$ and ε is an indep. centered noise s. t. $|\varepsilon| \le 1$.

Adaptive inference for non-parametric regression

Problem :

Non-parametric regression



Inference (estimation + uncertainty quantification) of the function?

The Model

f: function on $[0,1]^d$.

n observed data samples $(X_i, Y_i)_{i \leq n}$:

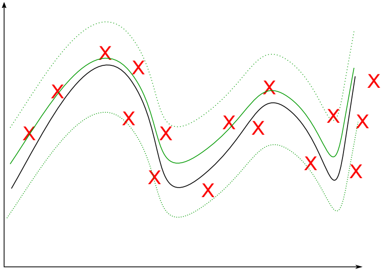
$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{[0,1]^d}$ and ε is an indep. centered noise s. t. $|\varepsilon| \leq 1$.

Adaptive inference for non-parametric regression

Problem :

Non-parametric regression



Inference (estimation + uncertainty quantification) of the function?

The Model

f: function on $[0,1]^d$.
n observed data samples $(X_i, Y_i)_{i \leq n}$:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{[0,1]^d}$ and ε is an indep. centered noise s. t. $|\varepsilon| \leq 1$.

$$\mathcal{C}(\alpha) = \quad \{\text{Hoelder ball} \quad (\alpha)\}.$$

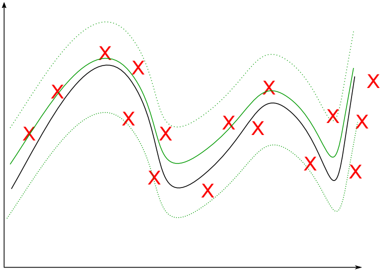
E.g. for $\alpha \leq 1$

$$\{f : |f(x) - f(y)| \leq \|x - y\|_\infty^\alpha\}.$$

Adaptive inference for non-parametric regression

Problem :

Non-parametric regression



Inference (estimation + uncertainty quantification) of the function?

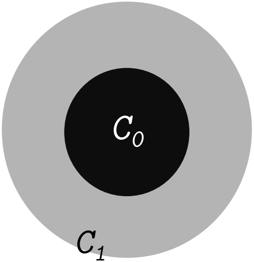
Question : If $f \in \mathcal{C}(\alpha)$, then the "optimal" precision of inference should depend on α. Inference adaptive to α?

Adaptive inference

Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive inference : Adaptation to the set $\mathcal{C}_h$ when $f \in \mathcal{C}_h$, $h \in \{0, 1\}$.



Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm

Minimax-opt. est. error

$$r_h = \inf_{\tilde{f} \text{ est.}} \sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\tilde{f} - f\|, \ \ h \in \{0, 1\}.$$

Minimax-optimal $\|.\|_\infty$ est. error in non-param. reg. $\mathcal{C}(\alpha)$:

$$\square \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)}.$$

See [Lepski, 1990-92, etc].

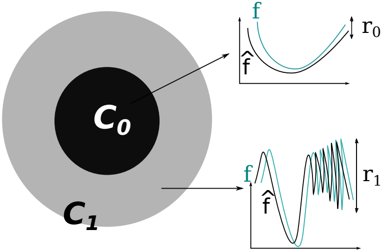
Adaptive inference

Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
  e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm



Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm
- In many models : adaptive estimator $\hat{f}$ exists

Adaptive estimation

$$\sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\hat{f} - f\| \leq \Box r_h, \quad \forall h \in \{0, 1\}.$$

Adaptive estimators exist in non-param. reg. See [Lepski, 1990-92, Donoho and Johnstone, 1998, etc].

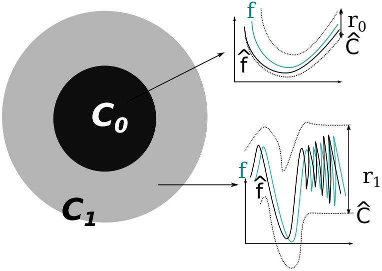
Adaptive inference

Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter



Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
  e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter

η-adapt. and honest conf. set

Honesty :

$$\sup_{f \in \mathcal{C}_1} \mathbb{P}_f(f \in \hat{C}) \geq 1 - \eta.$$

Adaptivity :

$$\sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\hat{C}\| \leq \square r_h, \quad \forall h \in \{0, 1\}.$$

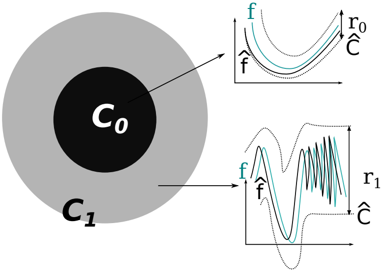
Adaptive inference

Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter



Adaptive inference

Adaptive estimation and confidence statements : See [Lepski, 1990-92], [Juditsky and Lambert-Lacroix, 1994], [Donoho and Johnstone, 1990-92], [Low, 2004-06], [Birgé and Massart, 1994-00], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
  e.g. $\mathcal{C}_0 =: \mathcal{C}(\gamma)$ and $\mathcal{C}_1 =: \mathcal{C}(\alpha)$ with $\alpha < \gamma$.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

In non-parametric regression : Adaptive and honest confidence sets do not exist. See [Cai and Low (2004)], [Hoffmann and Nickl (2011)], etc.

Indeed minimax rate for testing between $\mathcal{C}_0 = \mathcal{C}(\gamma)$ and $\mathcal{C}_1 = \mathcal{C}(\alpha)$ in $\|.\|_\infty$ norm is:

$$\left(\frac{\log(n)}{n}\right)^{-\alpha/(2\alpha+d)} = r_1 \gg r_0.$$

Common situation, adaptive inference paradox - see [Gine and Nickl, 2011], [C, Klopp, Löffler, Nickl, 2017] for a systematic study and relations to a testing problem.

Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Alice} & \text{Bob} & \text{Carine} & \text{Daniel} & \text{Ed} \end{array} \\
\begin{bmatrix}
\frownie & & \smiley & & \\
 & \frownie & & \smiley & \smiley \\
 & & \frownie & \frownie & \\
\smiley & \smiley & & & \frownie \\
 & & \smiley & \smiley &
\end{bmatrix}
\end{array}
$$

Inference (estimation + uncertainty quantification) of the matrix?

Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

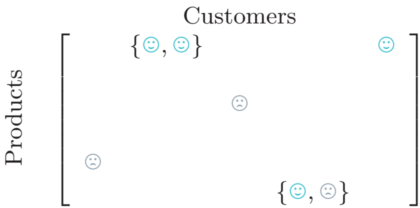
Inference (estimation + uncertainty quantification) of the matrix?

Trace Regression Model

f : matrix of dimension $d \times d$.
n observed data samples $(X_i, Y_i)_{i \leq n}$:

$$Y_i = f_{X_i} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{\{1,\dots,d\}^2}$ and ε is an indep. centered noise s. t. $|\varepsilon| \leq 1$.



Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

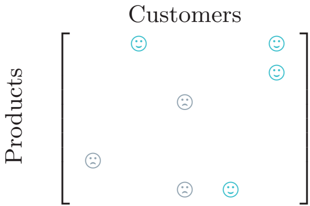
Bernoulli Model

f : matrix of dimension $d \times d$.

Data

$$Y_{i,j} = (f_{i,j} + \varepsilon_{i,j}) B_{i,j}, \ \ (i,j) \in \{1, \dots, d\}^2,$$

where $B_{i,j} \sim_{iid} \mathcal{B}(n/d^2)$ and ε is an indep. centered noise such that $|\varepsilon| \leq 1$.



Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

High dimensional regime : $d^2 \geq n$.

Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

$$
\begin{array}{ccccc}
\text{Alice} & \text{Bob} & \text{Carine} & \text{Daniel} & \text{Ed} \\
\end{array}
$$

$$
\begin{bmatrix}
\text{☹} & & \text{☺} & & \\
 & \text{☹} & & \text{☺} & \text{☺} \\
 & & \text{☹} & \text{☹} & \\
\text{☺} & \text{☺} & & & \text{☹} \\
 & & \text{☺} & \text{☺} &
\end{bmatrix}
$$

Inference (estimation + uncertainty quantification) of the matrix?

Let for $1 \le k \le d$

$$\mathcal{C}(k) = \{f : \mathrm{rank}(f) \le k, \|f\|_\infty \le 1\}.$$

High dimensional regime : $d^2 \ge n$.

Subtle problem : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

$$\begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Alice} & \text{Bob} & \text{Carine} & \text{Daniel} & \text{Ed} \end{array} \\ \left[\begin{array}{ccccc} \text{☹} & & \text{☺} & & \\ & \text{☹} & & \text{☺} & \text{☺} \\ & & \text{☹} & \text{☹} & \\ \text{☺} & \text{☺} & & & \text{☹} \\ & & \text{☺} & \text{☺} & \end{array}\right] \end{array}$$

Inference (estimation + uncertainty quantification) of the matrix?

High dimensional regime : $d^2 \geq n$.

Let for $1 \leq k \leq d$

$$\mathcal{C}(k) = \{f : \mathrm{rank}(f) \leq k, \|f\|_\infty \leq 1\}.$$

Question : If $f \in \mathcal{C}(k)$, then the "optimal" precision of inference should depend on k. Inference adaptive to k?

Adaptive estimation

There exists an adaptive estimator $\hat{f}$ of $f \in \mathcal{C}(k)$ that achieves the minimax-optimal error r_k over all $\mathcal{C}(k)$

$$\mathbb{E}\|\tilde{f} - f\|_F \leq \Box d\sqrt{\frac{kd}{n}} := \Box r_k.$$

where $\|.\|$ is the Frobenius norm, [Keshavan et al., 2009, Cai et al., 2010, Kolchinskii et al., 2011, Klopp and Gaiffas, 2015].

In terms of *estimation* of f, these two models are equivalent.

Question : Adaptive and honest confidence set scaling with r_k?

Adaptive estimation

There exists an adaptive estimator $\hat{f}$ of $f \in \mathcal{C}(k)$ that achieves the minimax-optimal error r_k over all $\mathcal{C}(k)$

$$\mathbb{E}\|\tilde{f} - f\|_F \leq \Box d\sqrt{\frac{kd}{n}} := \Box r_k.$$

where $\|.\|$ is the Frobenius norm, [Keshavan et al., 2009, Cai et al., 2010, Kolchinskii et al., 2011, Klopp and Gaiffas, 2015].

Question : Adaptive and honest confidence set scaling with r_k?

Matrix completion : Trace regression

Problem :

Application : Recommendation system (e.g. Netflix).

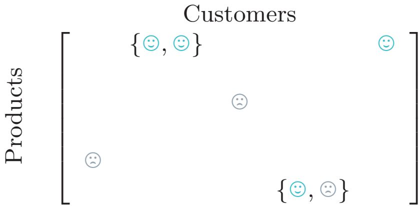
Inference (estimation + uncertainty quantification) of the matrix?

Trace Regression Model

f : matrix of dimension $d \times d$.
n observed data samples $(X_i, Y_i)_{i \le n}$:

$$Y_i = f_{X_i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{\{1,\ldots,d\}^2}$ and ε is an indep. centered noise s. t. $|\varepsilon| \le 1$.



Confidence sets : Trace Regression Model

Theorem (C., Klopp, Löffler and Nickl, 2016)

In the matrix completion "trace regression" model, η-adaptive and honest confidence sets exist.

Dimension reduction in the smaller model not too radical.

Matrix completion : Bernoulli Model

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

Bernoulli Model

f : matrix of dimension $d \times d$.
Data

$$Y_{i,j} = (f_{i,j} + \varepsilon_{i,j}) B_{i,j}, \ \ (i,j) \in \{1, \dots, d\}^2,$$

where $B_{i,j} \sim_{iid} \mathcal{B}(n/d^2)$ and ε is an indep. centered noise such that $|\varepsilon| \leq 1$.

Confidence sets : Bernoulli Model

Theorem (C., Klopp, Löffler and Nickl, 2016)

- **Bernoulli Model with known noise variance** *:* *Adaptive and honest confidence sets exist.*
- **Bernoulli Model with unknown noise variance** *:* *Adaptive and honest confidence sets do not exist .*

The two models are not equivalent in this case!

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} \smiley & \smiley & \frownie & \smiley & \smiley \\ \smiley & \frownie & \smiley & \smiley & \smiley \\ \smiley & \smiley & \frownie & \smiley & \frownie \\ \smiley & \frownie & \smiley & \smiley & \frownie \\ \frownie & \frownie & \smiley & \frownie & \smiley \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} \smiley & \smiley & \frownie & \smiley & \frownie \\ \frownie & \frownie & \smiley & \frownie & \smiley \\ \smiley & \smiley & \frownie & \smiley & \frownie \\ \smiley & \smiley & \frownie & \smiley & \frownie \\ \frownie & \frownie & \smiley & \frownie & \smiley \end{bmatrix} \end{array}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \text{☹} & & \text{☺} \\ & & \text{☺} & \text{☺} \\ & \text{☹} & & \text{☹} \\ \text{☺} & & & \\ \text{☹} & & & \text{☺} \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \text{☹} & & \text{☹} \\ & & \text{☹} & \text{☺} \\ & \text{☹} & & \text{☹} \\ \text{☺} & & & \\ \text{☹} & & & \text{☺} \end{bmatrix} \end{array}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \frownie & & \smiley \\ & & - & - \\ & \frownie & & \frownie \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \frownie & & \frownie \\ & & - & - \\ & \frownie & & \frownie \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & ☹ & & ☺ \\ & & - & - \\ & ☹ & & ☹ \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & ☹ & & ☹ \\ & & - & - \\ & ☹ & & ☹ \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

Less than $\frac{n^4}{d^4}$ such cycles whp $\rightarrow$ distinguishability only if $n \gg d$.

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!

Customers

$$\text{Products}\ \begin{bmatrix} ☺ & ☺ & ☹ & ☺ & ☺ \\ ☺ & ☹ & ☺ & ☺ & ☺ \\ ☺ & ☺ & ☹ & ☺ & ☹ \\ ☺ & ☹ & ☺ & ☺ & ☹ \\ ☹ & ☹ & ☺ & ☹ & ☺ \end{bmatrix}$$

H_1 : Rank one opinions.

Customers

$$\text{Products}\ \begin{bmatrix} ☺ & ☺ & ☹ & ☺ & ☺ \\ ☹ & ☹ & ☺ & ☹ & ☹ \\ ☺ & ☺ & ☹ & ☹ & ☹ \\ ☺ & ☺ & ☹ & ☹ & ☹ \\ ☹ & ☹ & ☺ & ☹ & ☹ \end{bmatrix}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & & \frownie & & \smiley \\ & & & \smiley & \smiley \\ & & \frownie & & \frownie \\ \smiley & & & & \\ \frownie & & & & \smiley \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & & \frownie & & \smiley \\ & & & \frownie & \frownie \\ & & \frownie & & \frownie \\ \smiley & & & & \\ \frownie & & & & \frownie \end{bmatrix} \end{array}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \frownie & & \smiley \\ & & - & - \\ & \frownie & & \frownie \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

H_1 : Rank one opinions.

$$\begin{array}{cc} & \text{Customers} \\ \text{Products} & \begin{bmatrix} & \frownie & & \smiley \\ & & - & - \\ & \frownie & & \frownie \\ - & & & \\ - & & & - \end{bmatrix} \end{array}$$

(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!

$$
\begin{array}{cc}
 & \text{Customers} \\
\text{Products} & \begin{bmatrix} & \text{☹} & & \text{☺} \\ & & - & - \\ & \text{☹} & & \text{☹} \\ - & & & \\ - & & & - \end{bmatrix}
\end{array}
$$

H_1 : Rank one opinions.

$$
\begin{array}{cc}
 & \text{Customers} \\
\text{Products} & \begin{bmatrix} & \text{☹} & & \text{☺} \\ & & - & - \\ & \text{☹} & & \text{☹} \\ - & & & \\ - & & & - \end{bmatrix}
\end{array}
$$

Less than $\frac{n^4}{d^4 k^3}$ *correct* cycles (taking rank groups into account)
$\rightarrow$ distinguishability only if $n \gg k^{3/4} d$.

Conclusion on adaptive inference

Adaptive inference paradox: adaptive estimation is generally possible and adaptive uncertainty quantification mostly not.

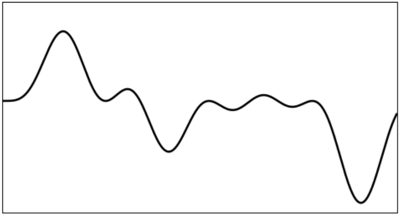
We have seen that in the non-parametric regression with L_∞ norm:

- Adaptive estimation is possible
- Adaptive and honest confidence sets do not exist

Typical example of adaptive inference paradox.

ADAPTIVITY IN $\mathcal{X}$ ARMED BANDITS

Non-Convex Optimization



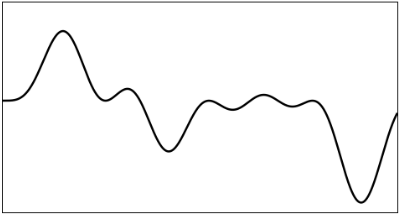
Problem

Finding/Exploiting the maximum $M(f)$ of an unknown function f.

Question

Can we design algorithms that adapt to the difficulty of the problem?

Non-Convex Optimization



Problem

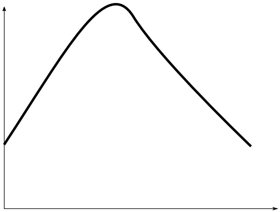
Finding/Exploiting the maximum $M(f)$ of an unknown function f.

Question

Can we design algorithms that adapt to the difficulty of the problem?

Non-Convex Optimization

Depending on the difficulty of the problem, we would hope to get different performances :



Question

Can we adapt to the hyperparameters?

$\mathcal{X}$-armed bandit problem

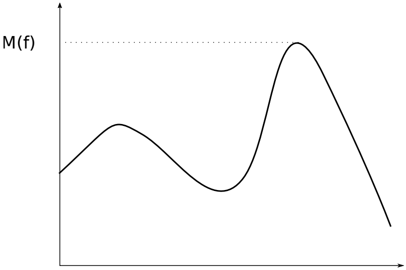
Game:

- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$

$\mathcal{X}$-armed bandit problem

Game:

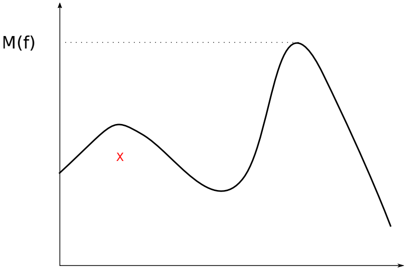
- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0, 1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0, 1]^d$



$\mathcal{X}$-armed bandit problem

Game:

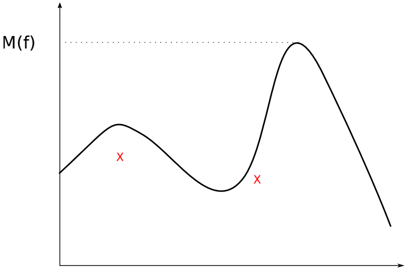
- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$



$\mathcal{X}$-armed bandit problem

Game:

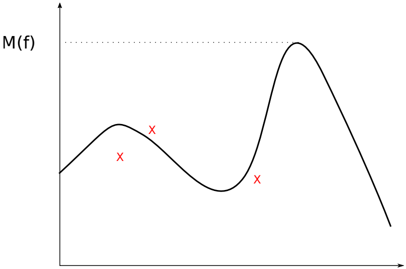
- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$



$\mathcal{X}$-armed bandit problem

Game:

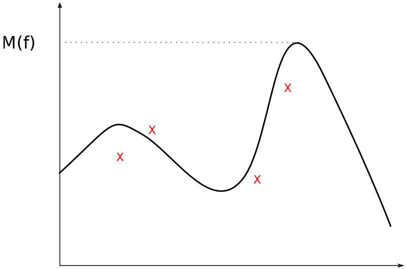
- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$



$\mathcal{X}$-armed bandit problem

Game:

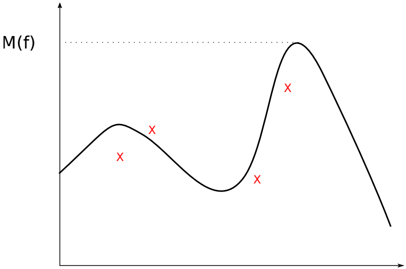
- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$



$\mathcal{X}$-armed bandit problem

Game:

- **Parameters :** function f with $M(f) = \max_x f(x)$, n
- **for** $t = 1, ..., n$
  - learner picks $X_t \in [0,1]^d$
  - receives $Y_t = f(X_t) + \epsilon_t$, ϵ indep. noise s.t. $\mathbb{E}\epsilon_t = 0, |\epsilon_t| \leq 1$
- **output** $x(n) \in [0,1]^d$



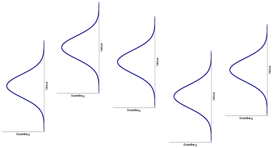
Performance measures:

- Simple regret: $r_n = M(f) - f(x(n))$
- Cumulative regret: $R_n = nM(f) - \sum_{t=1}^{n} f(X_t)$

Classical result for stochastic bandits

$K-$armed stochastic bandits

In the discrete case - f constant by parts on K known sets - classical stochastic bandits.



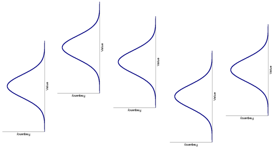
Idea

Approximate the continuous function f in K_n 'relevant' parts - this will depend on the regularity of f.

Classical result for stochastic bandits

$K-$armed stochastic bandits

In the discrete case - f constant by parts on K known sets - classical stochastic bandits.



The minimax regret satisfies (up to logarithmic terms)

$$\inf_{algo\ \mathcal{A}} \sup_{K-discrete\ f} r_n(\mathcal{A}, f) \approx \sqrt{\frac{K}{n}},$$

and

$$\inf_{algo\ \mathcal{A}} \sup_{K-discrete\ f} R_n(\mathcal{A}, f) \approx \sqrt{nK}.$$

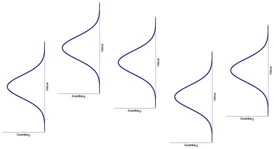
Idea

Approximate the continuous function f in K_n 'relevant' parts - this will depend on the regularity of f.

Classical result for stochastic bandits

$K-$armed stochastic bandits

In the discrete case - f constant by parts on K known sets - classical stochastic bandits.



The minimax regret satisfies (up to logarithmic terms)

$$\inf_{algo\ \mathcal{A}} \sup_{K-discrete\ f} r_n(\mathcal{A}, f) \approx \sqrt{\frac{K}{n}},$$

and

$$\inf_{algo\ \mathcal{A}} \sup_{K-discrete\ f} R_n(\mathcal{A}, f) \approx \sqrt{nK}.$$

Idea

Approximate the continuous function f in K_n 'relevant' parts - this will depend on the regularity of f.

Assumptions on f

Parametrize f from **easy** to **hard** problems.

Assumptions on f

Parametrize f from **easy** to **hard** problems.

- **Regularity condition:** $\exists \alpha > 0$ s.t. $\forall x, y$:

$$|f(x) - f(y)| \leq \|x - y\|_\infty^\alpha,$$

For $\alpha \leq 1$, Hölder assumption

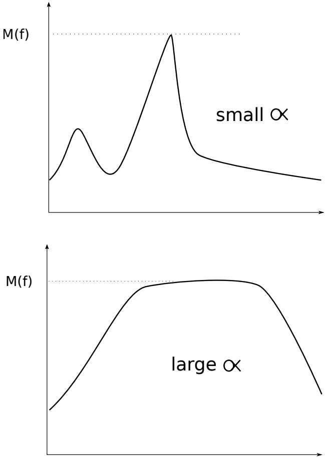
Assumptions on f

Parametrize f from **easy** to **hard** problems.

- **Regularity condition:** $\exists \alpha > 0$ s.t. $\forall x, y$:

$$|f(x) - f(y)| \leq \|x - y\|_\infty^\alpha,$$

For $\alpha \leq 1$, Hölder assumption



Assumptions on f

Parametrize f from **easy** to **hard** problems.

- **Regularity condition:** $\exists \alpha > 0$ s.t. $\forall x, y$:

$$|f(x) - f(y)| \leq \|x - y\|_\infty^\alpha,$$

For $\alpha \leq 1$, Hölder assumption

- **Margin condition:** $\exists \beta \geq 0$ s.t.:

$$\mathrm{Vol}(x : M(f) - f(x) \leq \Delta) \leq \Delta^\beta$$

No restriction for $\beta = 0$, larger β corresponds to easier problem

Assumptions on f

Parametrize f from **easy** to **hard** problems.

- **Regularity condition:** $\exists \alpha > 0$ s.t. $\forall x, y$:

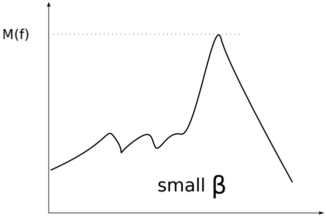
$$|f(x) - f(y)| \leq \|x - y\|_\infty^\alpha,$$

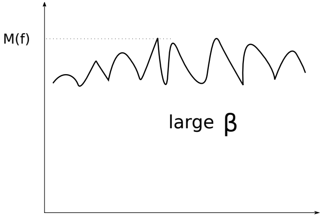
For $\alpha \leq 1$, Hölder assumption

- **Margin condition:** $\exists \beta \geq 0$ s.t.:

$$\mathrm{Vol}(x : M(f) - f(x) \leq \Delta) \leq \Delta^\beta$$

No restriction for $\beta = 0$, larger β corresponds to easier problem





Lower bounds

Define $\mathcal{P}(\alpha, \beta)$ the class of functions that satisfy these assumptions.

Theorem: Lower-bound for α, β known [Bubeck et al. 11]

For any strategy that performs at most n noisy function evaluations, it holds that:

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_P[r_n] \geq \square n^{-\frac{\alpha}{2\alpha+d-\alpha\beta}} := \square r_{\alpha,\beta},$$

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_P[R_n] \geq \square n^{1-\frac{\alpha}{2\alpha+d-\alpha\beta}} := \square R_{\alpha,\beta},$$

where $\square$ does not depend on the strategy and note that $R_{\alpha,\beta} = n r_{\alpha,\beta}$.

Goal: design procedures without access to α, β with optimal regret

Case α known

See e.g. [Agrawal (1995), Kleinberg (2004), Auer et al. (2007), Kleinberg et al. (2008), Bubeck et al. (2011a,b,c), Cope (2009), Munos (2014), Valko et al (2015)], etc.

Optimistic strategies (e.g. HOO in [Bubeck et al. 11]): use the knowledge of α to construct local (multi-scale) upper-confidence bounds on f, and choose next X_t optimistically.

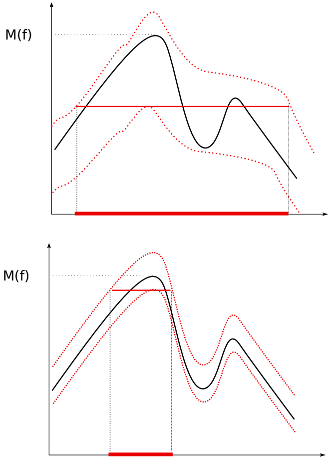
Our strategy: similar intuition but works hierarchically (at a single scale), only refining the partition in promising cells of the previous partition.

Case α known

See e.g. [Agrawal (1995), Kleinberg (2004), Auer et al. (2007), Kleinberg et al. (2008), Bubeck et al. (2011a,b,c), Cope (2009), Munos (2014), Valko et al (2015)], etc.

Optimistic strategies (e.g. HOO in [Bubeck et al. 11]): use the knowledge of α to construct local (multi-scale) upper-confidence bounds on f, and choose next X_t optimistically.

Our strategy: similar intuition but works hierarchically (at a single scale), only refining the partition in promising cells of the previous partition.



Case α known

Theorem: Upper-bound for α known [Locatelli, C, 2018]

Our strategy for opimisation is such that with probability at least $1 - n^{-1}$

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} r_n \leq \tilde{\square} r_{\alpha,\beta}, \qquad \mathbb{E}_P R_n \leq \tilde{\square} R_{\alpha,\beta}.$$

See also [Bubeck et al (2011), Minsker (2013), Bull (2014), Valko et al (2015)] etc.

Adaptivity?

The algorithm naturally adapts to β but needs α as a parameter.

Question

Can we adapt to α?

Adaptivity?

The algorithm naturally adapts to β but needs α as a parameter.

Question

Can we adapt to α?

Reminder from adaptive inference :

- Adaptive estimation is possible in non-parametric regression
- Adaptive and honest confidence sets do not exist in non-parametric regression

Adaptivity?

The algorithm naturally adapts to β but needs α as a parameter.

Question

Can we adapt to α?

Reminder from adaptive inference :

- Adaptive estimation is possible in non-parametric regression
- Adaptive and honest confidence sets do not exist in non-parametric regression

Question

Is the $\mathcal{X}$-armed bandit problem closer to adaptive estimation or adaptive and honest uncertainty quantification?

Adaptivity for simple regret (optimisation)

α-Adaptive strategy:

- Split budget in $\log^2 n$ chunks of same size
- Run previous strategy with $\alpha_i = \frac{i}{\log n}$ for all i
- Aggregate recommendations

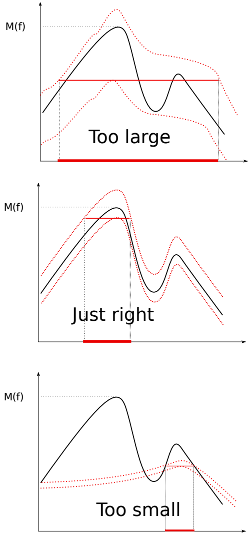
Idea: $\exists \alpha_{i^*}$ s.t.: $\alpha - \frac{1}{\log(n)} \leq \alpha_{i^*} \leq \alpha$

Adaptivity for simple regret (optimisation)

α-Adaptive strategy:

- Split budget in $\log^2 n$ chunks of same size
- Run previous strategy with $\alpha_i = \frac{i}{\log n}$ for all i
- Aggregate recommendations

Idea: $\exists \alpha_{i^*}$ s.t.: $\alpha - \frac{1}{\log(n)} \leq \alpha_{i^*} \leq \alpha$



Adaptivity for simple regret (optimisation)

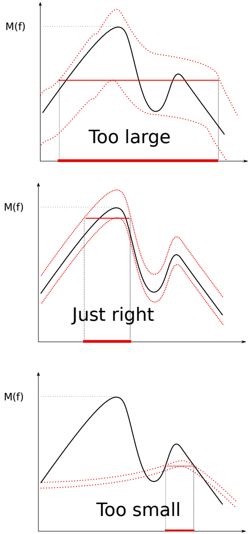
α-Adaptive strategy:

- Split budget in $\log^2 n$ chunks of same size
- Run previous strategy with $\alpha_i = \frac{i}{\log n}$ for all i
- Aggregate recommendations

Idea: $\exists \alpha_{i^*}$ s.t.: $\alpha - \frac{1}{\log(n)} \leq \alpha_{i^*} \leq \alpha$

- **Strategy 1:** cross-validate [Grill et al. 15]
- **Strategy 2:** recommend $x(n) \in \bigcap_{i \leq \hat{I}} s_{n,\alpha_i}$ [LCK 17]

Recovers the optimal rate but adaptively!



Adaptivity for simple regret (optimisation)

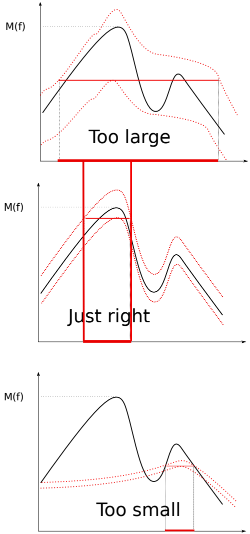
α-Adaptive strategy:

- Split budget in $\log^2 n$ chunks of same size
- Run previous strategy with $\alpha_i = \frac{i}{\log n}$ for all i
- Aggregate recommendations

Idea: $\exists \alpha_{i^*}$ s.t.: $\alpha - \frac{1}{\log(n)} \leq \alpha_{i^*} \leq \alpha$

- **Strategy 1:** cross-validate [Grill et al. 15]
- **Strategy 2:** recommend $x(n) \in \bigcap_{i \leq \hat{I}} s_{n,\alpha_i}$ [LCK 17]

Recovers the optimal rate but adaptively!



Adaptivity for simple regret (optimisation)

α-Adaptive strategy:

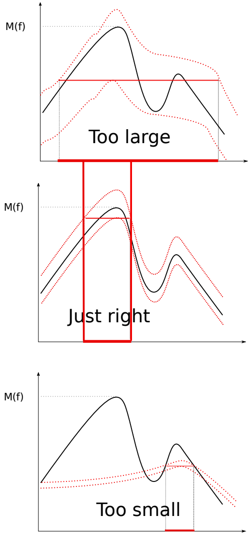
- Split budget in $\log^2 n$ chunks of same size
- Run previous strategy with $\alpha_i = \frac{i}{\log n}$ for all i
- Aggregate recommendations

Idea: $\exists \alpha_{i^*}$ s.t.: $\alpha - \frac{1}{\log(n)} \leq \alpha_{i^*} \leq \alpha$

Theorem: Upper-bound simple regret [Locatelli, C, 2018]

Our strategy yields whp

$$\sup_{\alpha,\beta \in S} \sup_{P \in \mathcal{P}(\alpha,\beta)} \frac{r_n}{\log(n)^u r_{\alpha,\beta}} \leq \square.$$



Adaptivity for Cumulative regret

Intuition: previous strategy favors exploration (linear regret)

Can we adaptively balance exploration/exploitation?

Theorem: Impossibility result for adaptive cumulative regret [Locatelli, C, 2018]

Fix $\gamma > \alpha > 0$ and β. Any strategy with (near)-optimal regret bounded by $\tilde{\Box} R_{\alpha,\beta}$ uniformly over $\mathcal{P}(\alpha, \beta)$ is such that

$$\sup_{P \in \mathcal{P}(\gamma,\beta)} \mathbb{E}_P[R_n] \geq \tilde{\Box} R_{\alpha,\beta}.$$

In fact something more refined holds for any algorithm with given rate on $\mathcal{P}(\gamma, \beta)$ or $\mathcal{P}(\alpha, \beta)$.

Adaptivity for Cumulative regret

Intuition: previous strategy favors exploration (linear regret)

Can we adaptively balance exploration/exploitation?

Theorem: Impossibility result for adaptive cumulative regret [Locatelli, C, 2018]

Fix $\gamma > \alpha > 0$ and β. Any strategy with (near)-optimal regret bounded by $\tilde{\Box} R_{\alpha,\beta}$ uniformly over $\mathcal{P}(\alpha, \beta)$ is such that

$$\sup_{P \in \mathcal{P}(\gamma,\beta)} \mathbb{E}_P[R_n] \geq \tilde{\Box} R_{\alpha,\beta}.$$

In fact something more refined holds for any algorithm with given rate on $\mathcal{P}(\gamma, \beta)$ or $\mathcal{P}(\alpha, \beta)$.

Adaptivity for Cumulative regret

Intuition: previous strategy favors exploration (linear regret)

Can we adaptively balance exploration/exploitation?

Theorem: Impossibility result for adaptive cumulative regret [Locatelli, C, 2018]

Fix $\gamma > \alpha > 0$ and β. Any strategy with (near)-optimal regret bounded by $\tilde{\square} R_{\alpha,\beta}$ uniformly over $\mathcal{P}(\alpha, \beta)$ is such that

$$\sup_{P \in \mathcal{P}(\gamma,\beta)} \mathbb{E}_P[R_n] \geq \tilde{\square} R_{\alpha,\beta}.$$

In fact something more refined holds for any algorithm with given rate on $\mathcal{P}(\gamma, \beta)$ or $\mathcal{P}(\alpha, \beta)$.

Conclusion

Adaptivity possible for simple regret but not for cumulative regret.

- Simple regret is in essence closer to adaptive estimation : adaptation possible
- Cumulative regret is in essence closer to adaptive and honest confidence sets : adaptation impossible

More systematic relation to adaptivity in active learning?