Adaptive Boosting in Online Non-Parametric Regression

55e Journées des Statistiques, Bordeaux

Paul Liautaud

May 28, 2024

Sorbonne University, Paris

Joint work with

Pierre Gaillard
CR Inria/UGA

Olivier Wintenberger
PR Sorbonne University

Table of Contents

Online Learning & Non-Parametric Regression

Setting & Problem

Data arrives **sequentially** as a stream

$$(x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (x_t, ?) \in \mathcal{X} \times \mathbb{R}$$

and we want to predict each next response y_t as as a function of x_t

$$\hat{f}_t(x_t)\,, \qquad \text{with } \hat{f}_t \in \mathbb{R}^{\mathcal{X}} \text{ sequentially updated.}$$

The scenario is as follows:

At each round $t = 1, \dots, T$, the learner or algorithm

1. observes input $x_t \in \mathcal{X}$
2. makes prediction $\hat{f}_t(x_t) \in \mathbb{R}$
3. incurs loss $\ell_t(\hat{f}_t(x_t))$
4. updates prediction function $\hat{f}_t \to \hat{f}_{t+1}$

Setting & Problem

At each round $t = 1, \dots, T$, the learner or algorithm

1. observes input $x_t \in \mathcal{X}$
2. makes prediction $\hat{f}_t(x_t) \in \mathbb{R}$ — Choose $\hat{f}_t$ before observing ℓ_t
3. incurs loss $\ell_t(\hat{f}_t(x_t))$ — No assumptions on how ℓ_t is generated!
4. updates prediction function $\hat{f}_t \to \hat{f}_{t+1}$

Setting & Notations

- $\ell_1, \dots, \ell_T$ are convex, differentiable and G-Lipschitz, with $G > 0$;
- $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ and we denote for any $\mathcal{X}' \subseteq \mathcal{X}$, $|\mathcal{X}'| = \sup_{x,x' \in \mathcal{X}'} \|x - x'\|_\infty$.

Setting & Problem

At each round $t = 1, \ldots, T$, the learner or algorithm

1. observes input $x_t \in \mathcal{X}$
2. makes prediction $\hat{f}_t(x_t) \in \mathbb{R}$
3. incurs loss $\ell_t(\hat{f}_t(x_t))$
4. updates prediction function $\hat{f}_t \to \hat{f}_{t+1}$

Goal:

minimize the cumulative loss

$$\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t))$$

Setting & Problem

At each round $t = 1, \dots, T$, the learner or algorithm

1. observes input $x_t \in \mathcal{X}$
2. makes prediction $\hat{f}_t(x_t) \in \mathbb{R}$
3. incurs loss $\ell_t(\hat{f}_t(x_t))$
4. updates prediction function $\hat{f}_t \to \hat{f}_{t+1}$

Goal:

minimize the cumulative loss $\Leftrightarrow$ predict almost as well as the best function f^*

$$\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t)) \qquad \underbrace{\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t)) - \sum_{t=1}^{T} \ell_t(f^*(x_t))}_{:=\mathrm{Reg}_T(f^*)}$$

Difficulty: no stochastic assumption on data: arbitrary time-series!

Setting & Problem

***Non-Parametric* regression** means that we are interested in forecasters $(\hat{f}_t)$ whose regret

$$\mathrm{Reg}_T(f^*) = \underbrace{\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t))}_{\text{our performance}} \quad - \quad \underbrace{\sum_{t=1}^{T} \ell_t(f^*(x_t))}_{\text{reference performance}}$$

against benchmark functions $f^* \in \mathcal{F}$ (e.g., Lipschitz) is as **small** as possible.

Setting & Problem

***Non-Parametric* regression** means that we are interested in forecasters $(\hat{f}_t)$ whose regret

$$\mathrm{Reg}_T(f^*) = \underbrace{\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t))}_{\text{our performance}} \quad - \quad \underbrace{\sum_{t=1}^{T} \ell_t(f^*(x_t))}_{\text{reference performance}} \quad \underbrace{= o(T)}_{\text{goal}}$$

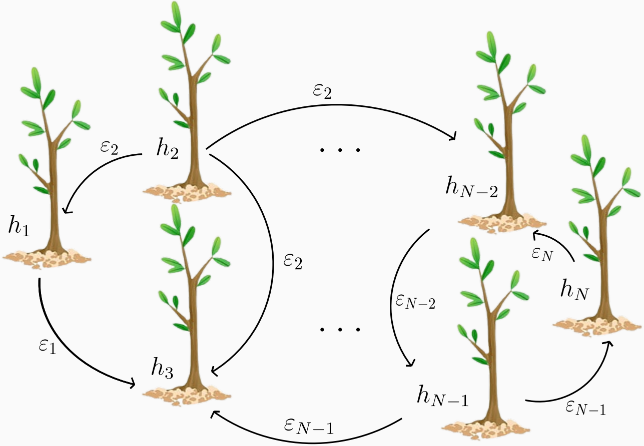
against benchmark functions $f^* \in \mathcal{F}$ (e.g., Lipschitz) is as **small** as possible.

Building Predictions with Online Gradient Boosting

Boosting uses ”wisdom of the crowd”

- **Boosting:** ensemble method combining multiple weak learners to create a strong learner
- Each model corrects/learns from errors of its peers

→ Resulting in a **highly accurate** predictive model [1]



[1] e.g. AdaBoost and XGBoost

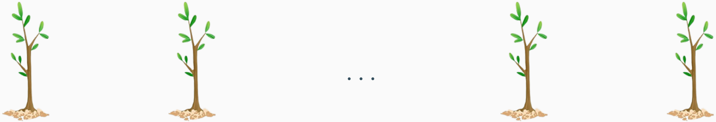
How to deal with weak learners?

- $\mathcal{W} \subset \mathbb{R}^{\mathcal{X}}$ a set of real valued functions $\mathcal{X} \to \mathbb{R}$;
- $\mathrm{span}_N(\mathcal{W}) = \{\sum_{n=1}^{N} \beta_n h_n,\, h_n \in \mathcal{W}, \beta_n \in \mathbb{R}\}$ linear function space associated to $\mathcal{W}$.

How to deal with weak learners?

- $\mathcal{W} \subset \mathbb{R}^{\mathcal{X}}$ a set of real valued functions $\mathcal{X} \to \mathbb{R}$;
- $\mathrm{span}_N(\mathcal{W}) = \{\sum_{n=1}^N \beta_n h_n,\ h_n \in \mathcal{W}, \beta_n \in \mathbb{R}\}$ linear function space associated to $\mathcal{W}$.

For each $t = 1, \ldots, T$, we use $N \geqslant 1$ *sequential* predictors from $\mathcal{W}$



$h_{1,t}$ $\quad$ $h_{2,t}$ $\quad \ldots \quad$ $h_{N-1,t}$ $\quad$ $h_{N,t}$

and we form *strong predictor* at any time $t \geqslant 1$ as

$$\hat{f}_t = \sum_{n=1}^{N} \beta_{n,t} h_{n,t}\,, \qquad \beta_{n,t} \in \mathbb{R}, n \in [N]$$

A new Online Gradient Boosting procedure

$\rightarrow$ **Goal:** We want to find a sequence of functions

$$\hat{f}_t = \sum_n \beta_{n,t} h_{n,t} \in \mathrm{span}_N(\mathcal{W})\,, \qquad 1 \leqslant t \leqslant T,$$

minimizing regret against $\mathcal{F} = \mathrm{span}_N(\mathcal{W})$.

A new Online Gradient Boosting procedure

$\rightarrow$ **Goal:** We want to find a sequence of functions

$$\hat{f}_t = \sum_n \beta_{n,t} h_{n,t} \in \operatorname{span}_N(\mathcal{W}), \qquad 1 \leqslant t \leqslant T,$$

minimizing regret against $\mathcal{F} = \operatorname{span}_N(\mathcal{W})$.

At $t \geqslant 1$, each $n \in [N]$ is boosted with OGB as:

---

1. Predict $\hat{f}_t(x_t)$,
2. $(\beta_{n,t}, h_{n,t})$ receives its gradient

$$g_{n,t} = \nabla_{(\beta_{n,t}, h_{n,t})} \ell_t(\hat{f}_t(x_t)),$$

3. Update as

$$(\beta_{n,t+1}, h_{n,t+1}) = \texttt{grad-step}((\beta_{n,t}, h_{n,t}), g_{n,t}). \tag{1}$$

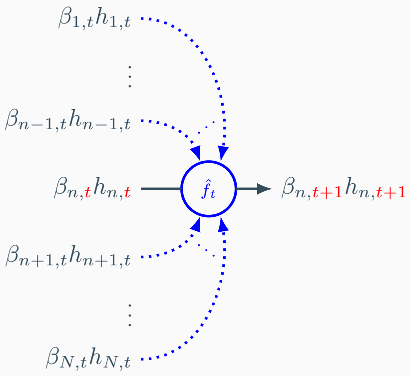


Figure 1: Boosting at time t.

Online Gradient Boosting in Chaining-Tree

Tree-based Method

Regular decision-tree $(\mathcal{T}, \bar{\mathcal{X}}, \bar{\mathcal{W}})$ over $\mathcal{X}$ is made of:

- a set of nodes $\mathcal{N}(\mathcal{T})$ including leaves $\mathcal{L}(\mathcal{T})$;
- a family of subregions

$$\bar{\mathcal{X}} = \{\mathcal{X}_n, n \in \mathcal{N}(\mathcal{T})\}$$

  partitionning $\mathcal{X}$ by level ;

- a family of prediction functions

$$\bar{\mathcal{W}} = \{h_n, n \in \mathcal{N}(\mathcal{T})\}.$$

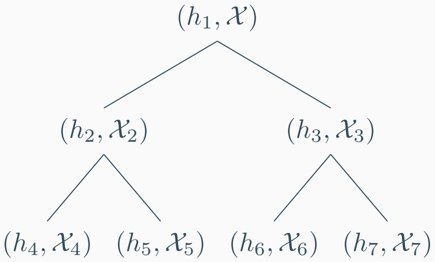


Figure 2: Example of $\mathcal{T}$ with depth $\mathrm{d}(\mathcal{T}) = 3$ over $\mathcal{X} \subset \mathbb{R}$.

Chaining-Tree

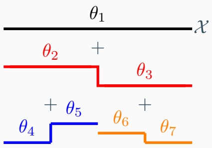


Figure 3: Prediction of a CT $\mathcal{T}$ of depth $\mathrm{d}(\mathcal{T}) = 3$ on $\mathcal{X} \subset \mathbb{R}$.

Definition (Chaining-Tree)

A Chaining-Tree (CT) prediction function $\hat{f}$ over $\mathcal{X}$ is defined as

$$\hat{f}(x) = \sum_{n \in \mathcal{N}(\mathcal{T})} h_n(x)\,, \quad x \in \mathcal{X},$$

where:

- $h_n(x) = \theta_n \mathbb{1}_{x \in \mathcal{X}_n}$ are constant functions;
- each interior node $n \in \mathcal{N}(\mathcal{T}) \setminus \mathcal{L}(\mathcal{T})$ has 2^d children forming a regular partition of $\mathcal{X}_n$.

Remark: contrary to standard methods, we predict with *all* nodes $n \in \mathcal{N}(\mathcal{T})$.

Illustration of approximation by Chaining-Tree

Assume ℓ_t is the square loss function and we launch a CT $\mathcal{T}$ with depth $\mathrm{d}(\mathcal{T}) = 1, 2, 3$, over T data. We have the following illustration:

Illustration of approximation by Chaining-Tree

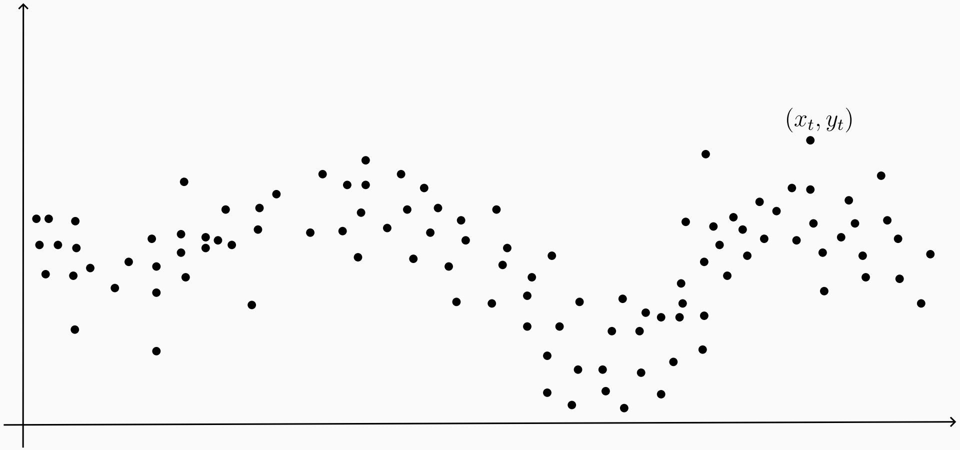


Illustration of approximation by Chaining-Tree

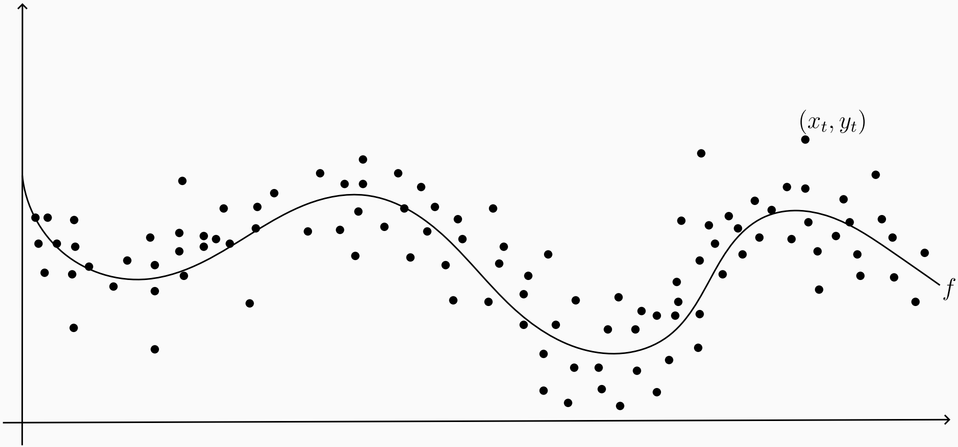


Illustration of approximation by Chaining-Tree

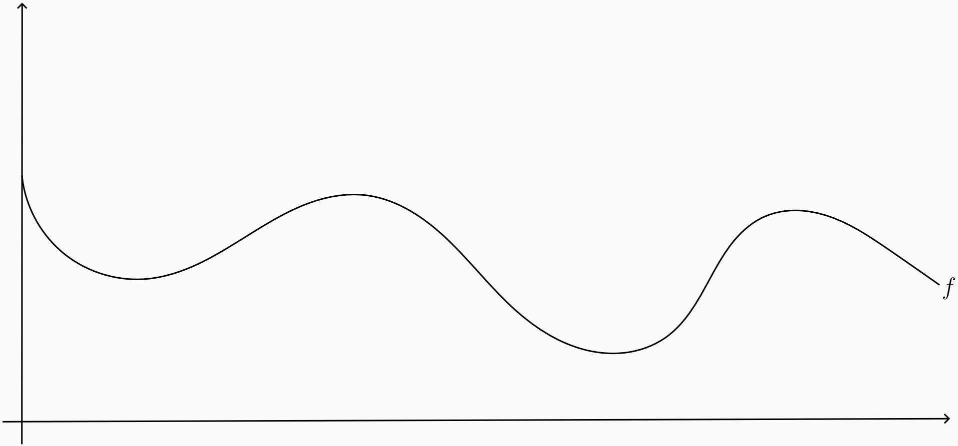


Illustration of approximation by Chaining-Tree

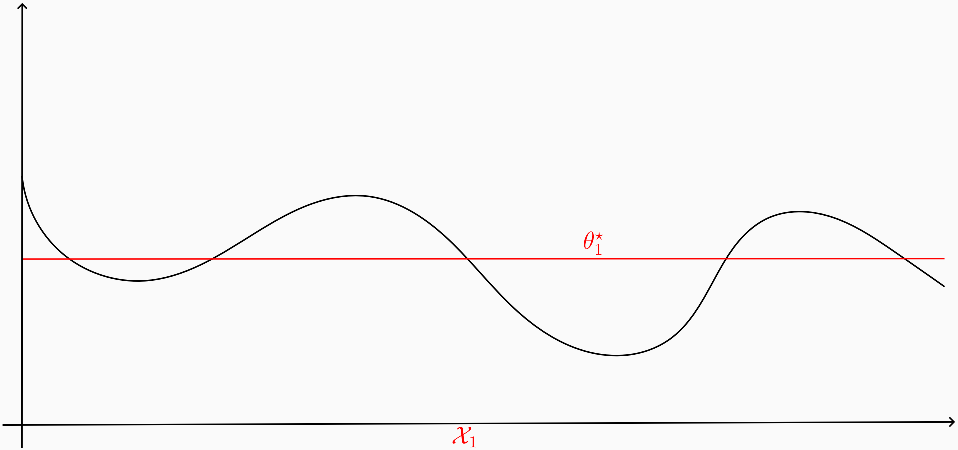


Illustration of approximation by Chaining-Tree

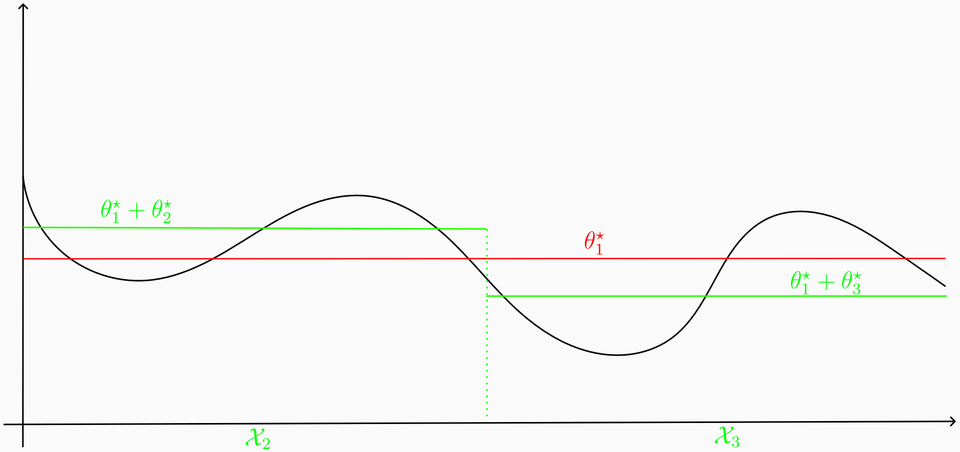
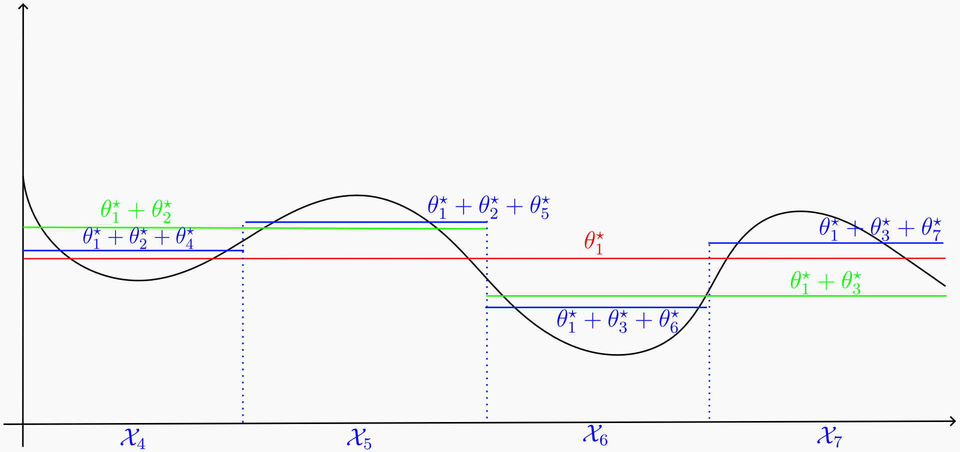


Illustration of approximation by Chaining-Tree



Online Boosting in a Chaining-Tree

→ **Goal:** Sequentially training CT $\mathcal{T}$, i.e. tuning over time the family

$$\bar{\mathcal{W}}_t = \{h_{n,t} = \theta_{n,t}\mathbb{1}_{\mathcal{X}_n}, n \in \mathcal{N}(\mathcal{T})\}.$$

We use OGB on $\bar{\mathcal{W}}_t$, with $\beta_n = 1$, $N = |\mathcal{N}(\mathcal{T})|$. Gradient step becomes, for all $n \in [N]$:

$$\theta_{n,t+1} \leftarrow \texttt{grad-step}(\theta_{n,t}, g_{n,t}), \quad \text{where} \quad g_{n,t} = \ell_t'(\hat{f}_t(x_t))\mathbb{1}_{x_t \in \mathcal{X}_n}.$$

Online Boosting in a Chaining-Tree

We use OGB on $\bar{\mathcal{W}}_t$, with $\beta_n = 1$, $N = |\mathcal{N}(\mathcal{T})|$. Gradient step becomes, for all $n \in [N]$:

$$\theta_{n,t+1} \leftarrow \texttt{grad-step}(\theta_{n,t}, g_{n,t}), \quad \text{where} \quad g_{n,t} = \ell'_t(\hat{f}_t(x_t))\mathbb{1}_{x_t \in \mathcal{X}_n}.$$

Which gradient step to consider? Any online optimization algorithm satisfying:

Assumtion 1

Let $n \in \mathcal{N}(\mathcal{T}), \forall g_{n,1}, \ldots, g_{n,T} \in [-G, G], G > 0$, parameters $(\theta_{n,t})$ satisfy:

$$\sum_{t \in T_n} g_{n,t}(\theta_{n,t} - \theta_n) \lesssim G|\theta_n|\sqrt{|T_n|}, \qquad \text{with} \quad T_n = \{1 \leqslant t \leqslant T, g_{n,t} \neq 0\},$$

for every $\theta_n \in \mathbb{R}$.

$\rightarrow$ *parameter free* algorithms (e.g. Cutkosky et al. (2018))

Optimal Regret and Adaptivity to Hölder functions

Hölder functions over $\mathcal{X} \subset \mathbb{R}^d$:

$$\mathrm{Lip}_L^\alpha(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} : |f(x) - f(x')| \leqslant L\|x - x'\|_\infty^\alpha\,, \forall x, x' \in \mathcal{X} \text{ and } \sup_{x \in \mathcal{X}} |f(x)| \leqslant L|\mathcal{X}|^\alpha\}.$$

Theorem (Regret of OGB-Chaining-Tree vs Hölder functions)

*Under Assumption 1, **OGB** on CT $(\mathcal{T}, \bar{\mathcal{X}}, \bar{\mathcal{W}})$ with $\mathcal{X}_{\mathrm{root}} = \mathcal{X}$, $\theta_{n,1} = 0, n \in \mathcal{N}(\mathcal{T})$ and $\mathrm{d}(\mathcal{T}) = \frac{1}{d}\log_2 T$ has regret:*

$$\sup_{f \in \mathrm{Lip}_L^\alpha(\mathcal{X})} \mathrm{Reg}_T(f) \lesssim GLX^\alpha \begin{cases} \sqrt{T} & \text{if } d < 2\alpha\,, \\ \log_2 T \sqrt{T} & \text{if } d = 2\alpha\,, \\ T^{1-\frac{\alpha}{d}} & \text{if } d > 2\alpha\,, \end{cases}$$

for any $L > 0, \alpha \in (0, 1]$.

Optimal Regret and Adaptivity to Hölder functions

Theorem (Regret of OGB-Chaining-Tree vs Hölder functions)

Under Assumption 1, ***OGB*** *on CT* $(\mathcal{T}, \bar{\mathcal{X}}, \bar{\mathcal{W}})$ *with* $\mathcal{X}_{\mathrm{root}} = \mathcal{X}$, $\theta_{n,1} = 0, n \in \mathcal{N}(\mathcal{T})$ *and* $\mathrm{d}(\mathcal{T}) = \frac{1}{d}\log_2 T$ *has regret:*

$$\sup_{f \in \mathrm{Lip}_L^\alpha(\mathcal{X})} \mathrm{Reg}_T(f) \lesssim GLX^\alpha \begin{cases} \sqrt{T} & \text{if } d < 2\alpha\,, \\ \log_2 T\sqrt{T} & \text{if } d = 2\alpha\,, \\ T^{1-\frac{\alpha}{d}} & \text{if } d > 2\alpha\,, \end{cases}$$

for any $L > 0, \alpha \in (0,1]$.

- Our rates are **minimax** over Lip_L^α (Rakhlin et al. (2015)) + we **do not need** prior knowledge of neither L nor α.
- Computationally tractable: x_t only falls into one subregion $\mathcal{X}_n$ for each level $1, \ldots, \mathrm{d}(\mathcal{T})$: we update $\mathcal{O}(\frac{T}{d}\log_2(T))$ for T rounds.

Adaptive Boosting in Online NonParametric Regression

Locally Adaptive Boosting - LocAdaBoost

We base our predictions on a core tree $\mathcal{T}_0$ as:

$$\hat{f}_t(x_t) = \sum_{n \in \mathcal{N}(\mathcal{T}_0)} w_{n,t} \hat{f}_{n,t}(x_t)\,, \quad \forall t \geqslant 1,$$

where for any $n \in \mathcal{N}(\mathcal{T}_0)$:

- $\hat{f}_n$ is a CT rooted at $\mathcal{X}_n$;
- $w_{n,t}$ weight associated.

We use OGB on

- $\beta_{n,t} = w_{n,t}$ with a specific `grad-step`;
- $\hat{f}_{n,t}$ as above.

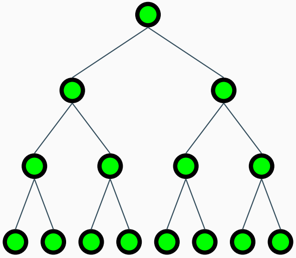
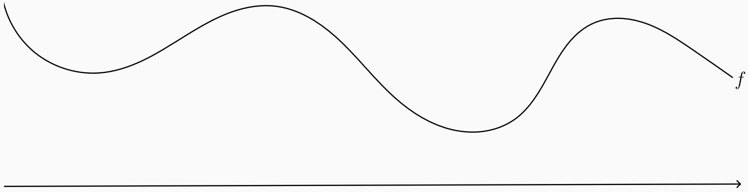


Figure 4: $\mathcal{T}_0$

Adaptivity to local profile of the competitor

→ **Goal:** Learn the best pruned tree from $\mathcal{T}_0$ in $\mathcal{P}(\mathcal{T}_0)$ to fit the competitor.

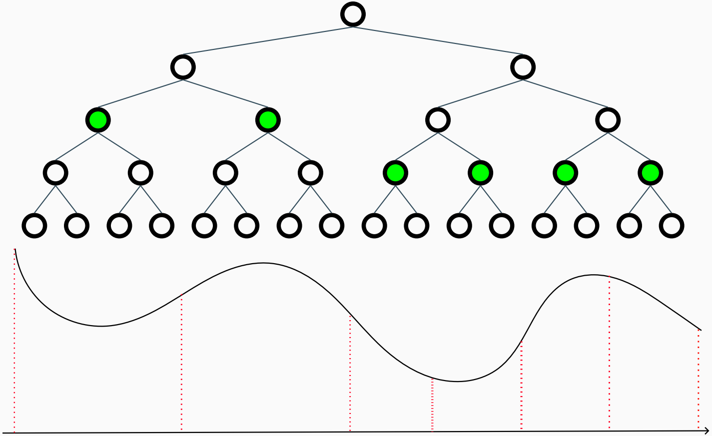
Example 1:



Adaptivity to local profile of the competitor

$\rightarrow$ **Goal:** Learn the best pruned tree from $\mathcal{T}_0$ in $\mathcal{P}(\mathcal{T}_0)$ to fit the competitor.

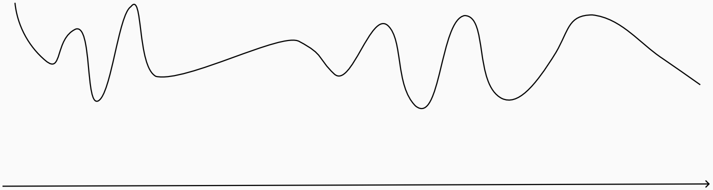
Example 1:



Adaptivity to local profile of the competitor

$\rightarrow$ **Goal:** Learn the best pruned tree from $\mathcal{T}_0$ in $\mathcal{P}(\mathcal{T}_0)$ to fit the competitor.

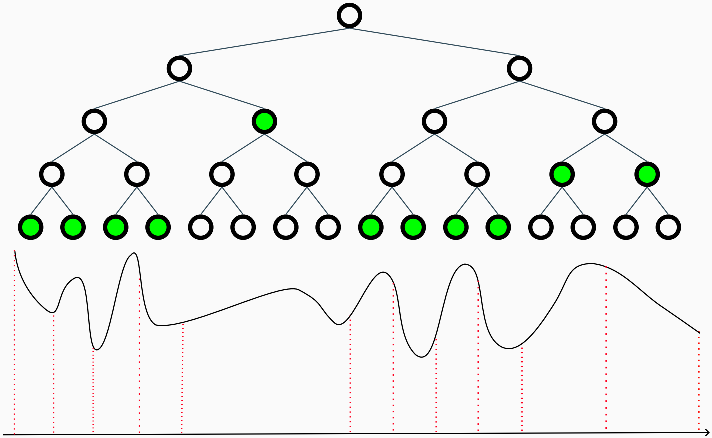
Example 2:



Adaptivity to local profile of the competitor

→ **Goal:** Learn the best pruned tree from $\mathcal{T}_0$ in $\mathcal{P}(\mathcal{T}_0)$ to fit the competitor.

Example 2:



Optimal and Locally Adaptive Regret (1/2)

Theorem (Locally Adaptive Regret, case $d = 1, \alpha > \frac{1}{2}$)

Under assumptions, for any $f \in \mathrm{Lip}_L^\alpha(\mathcal{X})$, `LocAdaBoost` *achieves*

$$\mathrm{Reg}_T(f) \lesssim \inf_{\mathcal{T} \in \mathcal{P}(\mathcal{T}_0)} \left\{ \sqrt{T|\mathcal{L}(\mathcal{T})|} + |\mathcal{L}(\mathcal{T})| + X^\alpha \sum_{n \in \mathcal{L}(\mathcal{T})} L_n(f) 2^{-\alpha \mathrm{d}(n)} \sqrt{|T_n|} \right\},$$

with $L_n(f)$ local Hölder constants.

If (ℓ_t) are exp-concave (e.g. square loss)

$$\mathrm{Reg}_T(f) \lesssim \inf_{\mathcal{T} \in \mathcal{P}(\mathcal{T}_0)} \left\{ |\mathcal{L}(\mathcal{T})| + X^\alpha \sum_{n \in \mathcal{L}(\mathcal{T})} L_n(f) 2^{-\alpha \mathrm{d}(n)} \sqrt{|T_n|} \right\}$$

Remark: `LocAdaBoost` could also adapt to local regularities (α_n)

Optimal and Locally Adaptive Regret (2/2)

Corollary (Minimax Regret)

For any $f \in \mathrm{Lip}_L^\alpha(\mathcal{X}), L > 0$, `LocAdaBoost` *achieves*

$$\mathrm{Reg}_T(f) \lesssim \begin{cases} \big(X^\alpha \bar{L}(f)\big)^{\frac{2}{2\alpha+1}} T^{\frac{1}{2\alpha+1}} & \textit{if } \ell_t \textit{ are exp-concave}, \\ \big(X^\alpha \bar{L}(f)\big)^{\frac{1}{2\alpha}} \sqrt{T}, \end{cases}$$

where $\bar{L}(f) = \big(\frac{1}{X} \sum_{n \in \mathcal{L}(\mathcal{T})} |\mathcal{X}_n| L_n(f)^{1/\alpha}\big)^\alpha$.

- ✓ Minimax optimality
- ✓ Adaptivity to local regularities (L_n) and α
- ✓ Adaptivity to the loss curvature

Conclusion

- New generic Online Gradient Boosting procedure;
- Online Gradient Boosting coupled with Chaining-Tree achieve minimax regret;
- Our unique `LocAdaBoost` algorithm both adapts optimaly to local regularities of the competitor and curvature of sequential losses;
- First constructive algorithm to achieve optimal locally adaptive regret;
- Future work: extend the boosting procedure to other learners to approach other classes of functions.

Thank you!

Questions?

References

Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Cutkosky, Ashok and Francesco Orabona (2018). "Black-box reductions for parameter-free online learning in banach spaces". In: *Conference On Learning Theory*. PMLR, pp. 1493–1529.

Gaillard, Pierre and Sebastien Gerchinovitz (2015). "A Chaining Algorithm for Online Nonparametric Regression". In: *COLT*.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2, pp. 169–192.

Mhammedi, Zakaria and Wouter M Koolen (2020). "Lipschitz and comparator-norm adaptivity in online learning". In: *Conference on Learning Theory*. PMLR, pp. 2858–2887.

Orabona, Francesco and Dávid Pál (2016). "Coin betting and parameter-free online learning". In: *Advances in Neural Information Processing Systems* 29.

Rakhlin, Alexander and Karthik Sridharan (2014). "Online non-parametric regression". In: *Conference on Learning Theory*. PMLR, pp. 1232–1264.

— (2015). "Online nonparametric regression with general loss functions". In: *arXiv preprint arXiv:1501.06598*.

Zinkevich, Martin (2003). "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936.