A chaining algorithm for online nonparametric regression

Sébastien Gerchinovitz

Institut de Mathématiques de Toulouse, Université Toulouse III – Paul Sabatier

This is joint work with Pierre Gaillard.
We consider the problem of online nonparametric regression. We present an algorithm that uses ideas based on the chaining technique.

Outline of the talk:

1. The chaining technique in the stochastic setting
2. Using chaining for online nonparametric regression
The chaining technique: a brief reminder

Technique introduced by Dudley (1967). Let \((X_f)_{f \in \mathcal{F}}\) be a centered stochastic process (indexed by a finite metric space \((\mathcal{F}, d)\)) with subgaussian increments:

\[
\forall f, g \in \mathcal{F}, \quad \forall \lambda > 0, \quad \log \mathbb{E} e^{\lambda(X_f - X_g)} \leq \frac{\lambda^2}{2} d(f, g)^2.
\]

**Goal:** upper bound the quantity \(\mathbb{E} \left[ \sup_{f \in \mathcal{F}} X_f \right] = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} (X_f - X_{f_0}) \right]\) for any \(f_0 \in \mathcal{F}\).

**Lemma (see, e.g., Boucheron et al. 2013)**

Let \(Z_1, \ldots, Z_K\) be such that \(\log \mathbb{E} e^{\lambda Z_i} \leq \lambda^2 \sqrt{\nu} / 2\) for all \(\lambda \in \mathbb{R}\).

Then, \(\mathbb{E} \max_{i=1, \ldots, K} Z_i \leq \sqrt{2\nu \log K}\).

This lemma entails the pessimistic bound (correlations are not used):

\(\mathbb{E} \left[ \sup_{f \in \mathcal{F}} (X_f - X_{f_0}) \right] \leq B \sqrt{2 \log (\text{card} \, \mathcal{F})}\) with \(B = \sup_{f \in \mathcal{F}} d(f, f_0)\).
Discretizing the space \((\mathcal{F}, d)\) into small balls

**Definition (metric entropy)**

- Let \((\mathcal{F}, d)\) be a metric space of finite cardinality.
- **\(\varepsilon\)-net**: any subset \(\mathcal{G} \subseteq \mathcal{F}\) such that
  \[
  \forall f \in \mathcal{F}, \exists g \in \mathcal{G} : d(f, g) \leq \varepsilon \iff \bigcup_{g \in \mathcal{G}} \bar{B}(g, \varepsilon) = \mathcal{F}
  \]

- \(\mathcal{N}_d(\mathcal{F}, \varepsilon)\): smallest cardinality of an \(\varepsilon\)-net.
- **metric entropy of \(\mathcal{F}\) at scale \(\varepsilon\)**: \(\log \mathcal{N}_d(\mathcal{F}, \varepsilon)\).
  It measures the complexity (richness) of the space \((\mathcal{F}, d)\).
Exploiting the correlations through the discretization

**Successive refining discretizations:**
Let $\mathcal{F}^{(0)} = \{f_0\}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(K-1)}, \mathcal{F}^{(K)} = \mathcal{F}$ be minimal $B/2^k$-nets of $\mathcal{F}$:

$$\forall f \in \mathcal{F}, \exists \pi_k(f) \in \mathcal{F}^{(k)}, d(f, \pi_k(f)) \leq B/2^k.$$

**Chaining argument:** using the lemma at multiple scales, we get:

$$\mathbb{E}\left[ \sup_{f \in \mathcal{F}} (X_f - X_{f_0}) \right] = \mathbb{E}\left[ \sup_{f \in \mathcal{F}} \sum_{k=1}^{K} \left( X_{\pi_k(f)} - X_{\pi_{k-1}(f)} \right) \right]$$

$$\leq \sum_{k=1}^{K} \mathbb{E}\left[ \sup_{f \in \mathcal{F}} \left( X_{\pi_k(f)} - X_{\pi_{k-1}(f)} \right) \right]$$

$$\leq 6 \sum_{k=1}^{K} B2^{-k} \sqrt{\log \mathcal{N}_d(\mathcal{F}, B/2^k)}$$

$$\leq 12B \int_{0}^{B/2} \sqrt{\log \mathcal{N}_d(\mathcal{F}, \varepsilon)} \, d\varepsilon.$$
Exploiting the correlations through the discretization

**Successive refining discretizations:**
Let $\mathcal{F}^{(0)} = \{f_0\}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(K-1)}, \mathcal{F}^{(K)} = \mathcal{F}$ be minimal $B/2^k$-nets of $\mathcal{F}$:

$$\forall f \in \mathcal{F}, \exists \pi_k(f) \in \mathcal{F}^{(k)}, d(f, \pi_k(f)) \leq B/2^k.$$

**Chaining argument:** using the lemma at multiple scales, we get:

$$\mathbb{E}\left[ \sup_{f \in \mathcal{F}} (X_f - X_{f_0}) \right] = \mathbb{E}\left[ \sup_{f \in \mathcal{F}} \sum_{k=1}^{K} (X_{\pi_k(f)} - X_{\pi_{k-1}(f)}) \right]$$

$$\leq \sum_{k=1}^{K} \mathbb{E}\left[ \sup_{f \in \mathcal{F}} \left( X_{\pi_k(f)} - X_{\pi_{k-1}(f)} \right) \right] \text{ small increments}$$

$$\leq 6 \sum_{k=1}^{K} B 2^{-k} \sqrt{\log \mathcal{N}_d(\mathcal{F}, B/2^k)}$$

$$\leq 12B \int_{0}^{B/2} \sqrt{\log \mathcal{N}_d(\mathcal{F}, \varepsilon)} d\varepsilon .$$

Dudley’s entropy integral
Setting: online regression with individual sequences

**Prediction task:** at each time $t \in \mathbb{N}^*$, predict the observation $y_t \in \mathbb{R}$ from the input $x_t \in \mathcal{X}$, on the basis of the past data $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$.

**Initial step:** the environment chooses arbitrary deterministic sequences $(y_t)_{t \geq 1}$ in $\mathbb{R}$ and $(x_t)_{t \geq 1}$ in $\mathcal{X}$ but the forecaster has not access to them.

**At each time round** $t \in \mathbb{N}^*$,

1. The environment reveals the input $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\hat{y}_t \in \mathbb{R}$.
3. The environment reveals the observation $y_t \in \mathbb{R}$ and the forecaster incurs the loss $(y_t - \hat{y}_t)^2$. 


Goal: minimizing regret

Let \( \mathcal{F} \subseteq \mathbb{R}^\mathcal{X} \) be a set of functions.

**Goal of the forecaster:** on the long run, to predict almost as well as the best function \( f \in \mathcal{F} \) in hindsight, that is, to minimize the regret:

\[
\text{Reg}_T(\mathcal{F}) \triangleq \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2.
\]

**Individual sequence setting:** our goal is to minimize the regret \( \text{Reg}_T(\mathcal{F}) \) uniformly over all sequences \( (y_t)_{t \geq 1} \) in \([-B, B]\) and \( (x_t)_{t \geq 1} \) in \( \mathcal{X} \); typically:

\[
\sup_{|y_t| \leq B, x_t \in \mathcal{X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} (y_t - f(x_t))^2 \right\} \leq o(1) \quad \text{when } T \to +\infty.
\]
Particular case: finite $F$

Assume that $F = \{f_1, f_2, \ldots, f_N\} \subseteq \mathbb{R}^X$ is finite. We can use a well known algorithm studied, e.g., by Kivinen and Warmuth (1999) and Vovk (2001):

**Algorithm (Exponentially Weighted Average forecaster (EWA))**

*Parameter: $\eta > 0$*

*At each round $t \geq 1$,*

- **Using past data, compute the weight vector** $\hat{\mathbf{w}}_t = (\hat{w}_{t,1}, \ldots, \hat{w}_{t,N})$ as

  $$\hat{w}_{t,j} \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (y_s - f_j(x_s))^2\right)}{\sum_{j'=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} (y_s - f_{j'}(x_s))^2\right)}, \quad 1 \leq j \leq N;$$

- **Compute the convex combination (convex aggregate):**

  $$\hat{y}_t \triangleq \sum_{j=1}^{N} \hat{w}_{t,j} f_j(x_t).$$
Regret guarantee when $\mathcal{F}$ is finite

If $\mathcal{F}$ contains $N$ functions, then we have a $O(\log N)$ upper bound on the regret under the boundedness assumption:

$$|y_1|, \ldots, |y_T| \leq B \quad \text{and} \quad \|f_1\|_{\infty}, \ldots, \|f_N\|_{\infty} \leq B .$$

**Theorem (Kivinen and Warmuth 1999)**

Assume that $\mathcal{F} = \{f_1, f_2, \ldots, f_N\} \subseteq [-B, B]^X$.

Then, the EWA algorithm tuned with $\eta = 1/(8B^2)$ satisfies: for all sequences $(y_t)_{t \geq 1}$ in $[-B, B]$ and $(x_t)_{t \geq 1}$ in $X$, for all $T \geq 1$,

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \min_{1 \leq j \leq N} \sum_{t=1}^{T} (y_t - f_j(x_t))^2 \leq 8B^2 \log N .$$

Remark 1: the requirement $\forall j, \|f_j\|_{\infty} \leq B$ can be removed.
Regret guarantee when $\mathcal{F}$ is finite

If $\mathcal{F}$ contains $N$ functions, then we have a $O(\log N)$ upper bound on the regret under the boundedness assumption:

$$|y_1|, \ldots, |y_T| \leq B \quad \text{and} \quad \|f_1\|_\infty, \ldots, \|f_N\|_\infty \leq B.$$

**Theorem (Kivinen and Warmuth 1999)**

Assume that $\mathcal{F} = \{f_1, f_2, \ldots, f_N\} \subseteq [-B, B]^X$.

Then, the EWA algorithm tuned with $\eta = 1/(8B^2)$ satisfies: for all sequences $(y_t)_{t \geq 1}$ in $[-B, B]$ and $(x_t)_{t \geq 1}$ in $X$, for all $T \geq 1$,

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \min_{1 \leq j \leq N} \sum_{t=1}^{T} (y_t - f_j(x_t))^2 \leq 8B^2 \log N.$$

Remark 1: the requirement $\forall j, \|f_j\|_\infty \leq B$ can be removed.
Remark 2: we can obtain a similar bound if $B = \max_{1 \leq t \leq T} |y_t|$ is unknown.
An extensive survey on prediction of individual sequences can be found in the following monograph:

Large function sets $\mathcal{F}$

Definition (metric entropy for sup norm)

- Let $\mathcal{F} \subseteq \mathbb{R}^X$ be a set of bounded functions endowed with the sup norm $\|f\|_\infty \triangleq \sup_{x \in X} |f(x)|$.

- **proper $\varepsilon$-net**: any subset $\mathcal{G} \subseteq \mathcal{F}$ such that
  \[ \forall f \in \mathcal{F}, \exists g \in \mathcal{G} : \|f - g\|_\infty \leq \varepsilon . \]

- $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$: smallest cardinality of a proper $\varepsilon$-net.

- **metric entropy of $\mathcal{F}$ at scale $\varepsilon$**: $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)$. 

Large function sets $\mathcal{F}$: suboptimal approach

Assume that $\mathcal{F}$ is infinite and, more precisely, that it is so large that
$log N_{\infty}(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p}$ as $\varepsilon \to 0$ (nonparametric, e.g., Hölder).

Natural approach (Vovk, 2006): approximate $\mathcal{F}$ with a minimal $\varepsilon$-net and run the EWA algorithm on this finite subset:

$$
\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \min_{1 \leq j \leq N_{\infty}(\mathcal{F}, \varepsilon)} \sum_{t=1}^{T} (y_t - f_j(x_t))^2 + 8B^2 \log N_{\infty}(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p}.
$$

Problem: when upper bounding $\min_{1 \leq j \leq N_{\infty}(\mathcal{F}, \varepsilon)}$ by $\min_{f \in \mathcal{F}}$, we loose an approximation term of order $\varepsilon T$. Optimizing in $\varepsilon$ only yields:

$$
\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 + O\left(\frac{T^p}{(p+1)}\right),
$$

which is suboptimal.
We still assume that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p}$ as $\varepsilon \to 0$.

**Optimal regret**: through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if $p \in (0, 2)$, then

$$\text{Reg}_T(\mathcal{F}) \leq c_1 B^2 (1 + \log \mathcal{N}_\infty(\mathcal{F}, \gamma)) + c_2 B \sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} d\varepsilon$$

$$\lesssim \gamma^{-p} + \sqrt{T} \int_0^\gamma \varepsilon^{-p/2} d\varepsilon$$

$$\lesssim T^{p/(p+2)} \quad \text{for} \quad \gamma = T^{-1/(p+2)}.$$

The rate $T^{p/(p+2)}$ is better than $T^{p/(p+1)}$ obtained previously with EWA, and it is (in a sense) optimal.
We still assume that \( \log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p} \) as \( \varepsilon \to 0 \).

**Optimal regret:** through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if \( p \in (0, 2) \), then

\[
\text{Reg}_T(\mathcal{F}) \leq c_1 B^2 \left( 1 + \log \mathcal{N}_\infty(\mathcal{F}, \gamma) \right) + c_2 B \sqrt{T} \int_0^{\gamma} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} d\varepsilon
\]

\[
\lesssim \gamma^{-p} + \sqrt{T} \int_0^{\gamma} \varepsilon^{-p/2} d\varepsilon
\]

\[
\lesssim T^{p/(p+2)} \quad \text{for} \quad \gamma = T^{-1/(p+2)}.
\]

**Example:** Hölder class \( \mathcal{F} \subseteq \mathbb{R}^{[0,1]} \) of regularity \( \beta = q + \alpha \):

\[
|f^{(q)}(x) - f^{(q)}(y)| \leq \lambda |x - y|^{\alpha} \quad \text{and} \quad \forall k \leq q, \|f^{(k)}\|_{\infty} \leq B
\]

\[
\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) = \mathcal{O}(\varepsilon^{-1/\beta}) \quad \Rightarrow \quad \text{Reg}_T(\mathcal{F})/T = \mathcal{O}(T^{-2\beta/(2\beta+1)}) \quad \text{if} \quad \beta > 1/2.
\]

\( \sim \) same rate as in the statistical setting (for \( \beta > 1/2 \)).
We still assume that \( \log N_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p} \) as \( \varepsilon \to 0 \).

**Optimal regret**: through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if \( p \in (0, 2) \), then

\[
\text{Reg}_T(\mathcal{F}) \leq c_1 B^2 (1 + \log N_\infty(\mathcal{F}, \gamma)) + c_2 B \sqrt{T} \int_0^\gamma \sqrt{\log N_\infty(\mathcal{F}, \varepsilon)} d\varepsilon
\]

\[
\lesssim \gamma^{-p} + \sqrt{T} \int_0^\gamma \varepsilon^{-p/2} d\varepsilon
\]

\[
\lesssim T^{p/(p+2)} \quad \text{for} \quad \gamma = T^{-1/(p+2)}.
\]

The above integral is a **Dudley entropy integral**. In statistical learning with i.i.d. data, it is useful to derive risk bounds for empirical risk minimizers (e.g., Massart 2007; Rakhlin et al. 2013).
Our contributions

1. We provide an explicit algorithm that achieves the Dudley-type regret bound (when $p \in (0, 2)$):

$$\text{Reg}_T(\mathcal{F}) \leq c_1 B^2 (1 + \log N_\infty(\mathcal{F}, \gamma)) + c_2 B \sqrt{T} \int_0^\gamma \sqrt{\log N_\infty(\mathcal{F}, \varepsilon)} d\varepsilon.$$

Nota: contrary to Rakhlin and Sridharan (2014), our bounds are not in terms of the stronger notion of "sequential entropy."

2. This algorithm uses ideas from the chaining technique, and relies on a new subroutine (Multi-variable Exponentiated Gradient algorithm) to perform optimization at different scales simultaneously.

3. We address computational issues by showing how to construct more efficient and quasi-optimal $\varepsilon$-nets.
Turning the chaining technique into an online algorithm

**Getting back to online nonparametric regression:** large class $\mathcal{F}$ satisfying $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p}$ as $\varepsilon \to 0$.

Recall that we want to prove a bound of the form:

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 + \text{[small term]}$$

**Chaining principle:** as previously, we discretize $\mathcal{F}$ and use projections $\pi_k(f)$ such that $\sup_f \|\pi_k(f) - f\|_\infty \leq \gamma/2^k$ for all $k \geq 0$.

$$\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 = \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(y_t - \pi_0(f)(x_t) - \sum_{k=1}^{\infty} \left[\pi_k(f) - \pi_{k-1}(f)\right](x_t)\right)^2$$

|small increments|$\leq 3\gamma/2^k$
Aggregation at two different levels

\[
\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 = \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left( y_t - \pi_0(f)(x_t) - \sum_{k=1}^{\infty} \left[ \pi_k(f) - \pi_{k-1}(f) \right](x_t) \right)^2
\]

Sufficient goal:
\[
\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \inf_{f_0, g_1, \ldots, g_K} \sum_{t=1}^{T} (y_t - (f_0 + g_1 + \ldots + g_K)(x_t))^2 + \text{[small term]}
\]

Two aggregation levels:

\[
\begin{align*}
  f_{0,1} & \xrightarrow{\text{low scale gradient descent}} \hat{f}_{t,1} \\
  f_{0,2} & \xrightarrow{} \hat{f}_{t,2} \\
  \vdots \\
  f_{0,N_0} & \xrightarrow{} \hat{f}_{t,N_0}
\end{align*}
\]

\[
\begin{align*}
  \hat{y}_t & = \sum_{j=1}^{N_0} \hat{w}_{t,j} \hat{f}_{t,j}(x_t) \\
  \text{EWA}
\end{align*}
\]
Combining two regret guarantees

**High-scale aggregation** Using an Exponentially Weighted Average (EWA) forecaster $\hat{f}_t = \sum_{j=1}^{N_0} \hat{w}_{t,j} \hat{f}_{t,j}$ yields

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \min_{1 \leq j \leq N_0} \sum_{t=1}^{T} \left( y_t - \hat{f}_{t,j}(x_t) \right)^2 + \Box B^2 \log N_0$$

**Low-scale aggregation** Recall that $G^{(k)} = \{\pi_k(f) - \pi_{k-1}(f) : f \in \mathcal{F}\}$. Denote $G^{(k)} = \{g_1^{(k)}, \ldots, g_{N_k}^{(k)}\}$.

We designed a multi-variable extension of the Exponentiated Gradient algorithm:

$$\hat{f}_{t,j} \triangleq f_{0,j} + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \hat{u}_{t,i}^{(j,k)} g_i^{(k)}$$

which yields, for all $j = 1, \ldots, N_0$,

$$\sum_{t=1}^{T} \left( y_t - \hat{f}_{t,j}(x_t) \right)^2 \leq \min_{g_1, \ldots, g_K} \sum_{t=1}^{T} \left( y_t - (f_{0,j} + g_1 + \ldots + g_K)(x_t) \right)^2$$

$$+ 120B \sqrt{T} \int_{0}^\gamma \sqrt{\log N_\infty(\mathcal{F}, \varepsilon)} d\varepsilon .$$
Main result

The next theorem indicates that the Chaining Exponentially Weighted Average forecaster satisfies a Dudley-type regret bound.

**Theorem (Gaillard and G., 2015)**

Let $B > 0$, $T \geq 1$, and $\gamma \in \left( \frac{B}{T}, B \right)$.

- Assume that $\max_{1 \leq t \leq T} |y_t| \leq B$ and that $\sup_{f \in F} \|f\|_\infty \leq B$.
- Assume that $(F, \|\cdot\|_\infty)$ is totally bounded and define $F^{(0)}$ and $G^{(k)}$ as above.

Then, the Chaining Exponentially Weighted Average forecaster (tuned with appropriate parameters) satisfies:

$$\text{Reg}_T(F) \leq B^2 \left( 5 + 50 \log N_\infty(F, \gamma) \right) + 120B\sqrt{T} \int_0^{\gamma/2} \sqrt{\log N_\infty(F, \varepsilon)} d\varepsilon.$$
We assume that \( \mathcal{F} = \{ f : [0, 1] \to [-B, B] : f \text{ is } 1\text{-Lipschitz} \} \).

**Regret bound:**
We know that \( \log N_\infty(\mathcal{F}, \varepsilon) = O(\varepsilon^{-1}) \).
Therefore, our algorithm obtains \( \text{Reg}_T(\mathcal{F}) = O(T^{1/3}) \), which is optimal.

**Computational issue:**
Our algorithm updates \( \exp(O(T)) \) weights at every round \( t \).
Hence very poor time and space computational complexities.

**Solution:**
\( \mathcal{F} \) has a sufficiently nice structure that can be exploited to construct computationally manageable \( \varepsilon \)-nets with quasi-optimal cardinality.

For example: histograms on a dyadic discretization lead to \( O(T^{1/3} \log T) \) regret and per-round time complexity.
Thank you for your attention!

For more details, the preprint is available on:
http://www.math.univ-toulouse.fr/~sgerchin/
Appendix
Lipschitz class $\mathcal{F}$: a computationally efficient discretization

We assume that $\mathcal{F} = \{ f : [0, 1] \to [-B, B] : f \text{ is } 1\text{-Lipschitz} \}$.

**Regret bound:**
We know that $\log N_{\infty}(\mathcal{F}, \varepsilon) = O(\varepsilon^{-1})$.
Therefore, our algorithm obtains $\text{Reg}_T(\mathcal{F}) = O(T^{1/3})$, which is optimal.

**Computational issue:**
Our algorithm updates exponentially many weights at every round $t$.
Hence poor time and space computational complexities.

**Solution:**
$\mathcal{F}$ has a sufficiently nice structure that can be exploited to construct computationally manageable $\varepsilon$-nets with quasi-optimal cardinality.
High-level discretization (piecewise-constant approximation)

- Partition the $x$-axis $[0,1]$: $I_a \triangleq [(a-1)\gamma, a\gamma)$, $a = 1, \ldots, \frac{1}{\gamma}$.
- Discretize the $y$-axis $[-B, B]$: $C^{(0)} = \{-B + j\gamma : j = 0, \ldots, \frac{2B}{\gamma}\}$.

$\mathcal{F}^{(0)}$: set of piecewise-constant functions $f^{(0)}(x) = \sum_{a=1}^{1/\gamma} c_a^{(0)} \mathbb{I}_{x \in I_a}$, $c_a^{(0)} \in C^{(0)}$. 

\[ 
\begin{array}{c}
\gamma \\
\vdots
\end{array} \\
\begin{array}{c}
\vdots \\
\vdots
\end{array} \\
\begin{array}{c}
c_1^{(0)} \\
\vdots \\
\vdots \\
\vdots \\
c_n^{(0)}
\end{array} \\
\begin{array}{c}
0 \\
\gamma \\
2\gamma \\
3\gamma \\
4\gamma \\
\vdots
\end{array}
\]
Low-level discretization (dyadic approximation)

\( \mathcal{F}^{(M)} \): set of all functions \( f_c : [0, 1] \rightarrow \mathbb{R} \) of the form

\[
f_c(x) = \sum_{a=1}^{1/\gamma} c_a^{(0)} \mathbb{I}_{x \in I_a} + \sum_{m=1}^{M} \sum_{a=1}^{1/\gamma} \sum_{n=1}^{2^m} c_a^{(m,n)} \mathbb{I}_{x \in I_a^{(m,n)}} .
\]
Theorem (Gaillard and G., 2015)

Let $B > 0$, $T \geq 2$, and $\mathcal{F}$ be the set of all 1-Lipschitz functions from $[0, 1]$ to $[-B, B]$. Assume that $\max_{1 \leq t \leq T} |y_t| \leq B$.

Then, the Dyadic Chaining Algorithm (see preprint) satisfies, for some absolute constant $c > 0$,

$$\text{Reg}_T(\mathcal{F}) \leq c \max\{B, B^2\} T^{1/3} \log T.$$  

Remark: additional log factor, but computationally tractable:

- per-round time complexity: $\mathcal{O}(T^{1/3} \log T)$;
- space complexity: $\mathcal{O}(T^{4/3} \log T)$.  


