Online Learning : Random Averages, Combinatorial Parameters and Learnability

Alexander Rakhlin,

Wharton School
University of Pennsylvania

Karthik Sridharan,

Toyota Technological Institute at Chicago

Ambuj Tewari

University of Texas at Austin

Statistical Vs Online Learning

Statistical Learning	Online Learning

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$
Certificate for learnability and rates : Complexity measures on $\mathcal{F}$	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$
Certificate for learnability and rates : Complexity measures on $\mathcal{F}$ Eg. : VC dimension, Rademacher Complexity, Covering Numbers, Fat-shattering Dimension ...	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$
Certificate for learnability and rates : Complexity measures on $\mathcal{F}$ Eg. : VC dimension, Rademacher Complexity, Covering Numbers, Fat-shattering Dimension ... Algorithm : Empirical Risk Minimization	

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$
Certificate for learnability and rates : Complexity measures on $\mathcal{F}$ Eg. : VC dimension, Rademacher Complexity, Covering Numbers, Fat-shattering Dimension ... Algorithm : Empirical Risk Minimization	Certificate for learnability and rates : Algorithmic : Algorithm + Regret Bound (case by case)

Statistical Vs Online Learning

Statistical Learning	Online Learning
Nature picks distribution D on $\mathcal{X}$ $\{x_t\}_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} D$ provided to learner Learner picks $\hat{f} \in \mathcal{F}$	For $t = 1$ to T Learner picks $f_t \in \mathcal{F}$ Adversary simultaneously picks $x_t \in \mathcal{X}$ End
Goal : $\mathbb{E}_D[\ell(\hat{f}, x)] - \inf_{f \in F} \mathbb{E}_D[\ell(f, x)] \to 0$	Goal : $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, x_t) - \inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, x_t) \to 0$
Certificate for learnability and rates : Complexity measures on $\mathcal{F}$ Eg. : VC dimension, Rademacher Complexity, Covering Numbers, Fat-shattering Dimension . . . Algorithm : Empirical Risk Minimization	Certificate for learnability and rates : Algorithmic : Algorithm + Regret Bound (case by case)

Complexity measures on $\mathcal{F}$ for Online Learning ?

Value of the Game

For $t = 1$ to T

Learner picks $f_t \in \mathcal{F}$

Adversary simultaneously picks $x_t \in \mathcal{X}$

Learner pays loss $\ell(f_t, x_t)$

End

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary simultaneously picks $x_t \in \mathcal{X}$

Learner pays loss $\ell(f_t, x_t)$

End

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

Learner pays loss $\ell(f_t, x_t)$

End

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $\ell(f_t, x_t)$

End

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t)\right]$$

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\inf_{q_1 \in \Delta(\mathcal{F})} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\inf_{q_1 \in \Delta(\mathcal{F})} \sup_{x_1 \in \mathcal{X}} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\inf_{q_1 \in \Delta(\mathcal{F})} \sup_{x_1 \in \mathcal{X}} \mathbb{E}_{f_1 \sim q_1} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\mathcal{V}_T(\mathcal{F}) := \inf_{q_1 \in \Delta(\mathcal{F})} \sup_{x_1 \in \mathcal{X}} \mathop{\mathbb{E}}_{f_1 \sim q_1} \cdots \inf_{q_T \in \Delta(\mathcal{F})} \sup_{x_T \in \mathcal{X}} \mathop{\mathbb{E}}_{f_T \sim q_T} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\mathcal{V}_T(\mathcal{F}) := \inf_{q_1 \in \Delta(\mathcal{F})} \sup_{x_1 \in \mathcal{X}} \mathop{\mathbb{E}}_{f_1 \sim q_1} \cdots \inf_{q_T \in \Delta(\mathcal{F})} \sup_{x_T \in \mathcal{X}} \mathop{\mathbb{E}}_{f_T \sim q_T} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Exists randomized online learning algorithm whose expected regret is bounded by $\mathcal{V}_T(\mathcal{F})$.

Value of the Game

For $t = 1$ to T

Learner picks $q_t \in \Delta(\mathcal{F})$

Adversary picks $x_t \in \mathcal{X}$

$f_t \sim q_t$ and learner pays loss $f_t(x_t)$

End

Regret of optimal learner against optimal adversary :

[Abernethy,Agarwal,Bartlett,Rakhlin'09]

$$\mathcal{V}_T(\mathcal{F}) := \inf_{q_1 \in \Delta(\mathcal{F})} \sup_{x_1 \in \mathcal{X}} \mathop{\mathbb{E}}_{f_1 \sim q_1} \cdots \inf_{q_T \in \Delta(\mathcal{F})} \sup_{x_T \in \mathcal{X}} \mathop{\mathbb{E}}_{f_T \sim q_T} \left[\sum_{t=1}^{T} f_t(x_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} f(x_t) \right]$$

Exists randomized online learning algorithm whose expected regret is bounded by $\mathcal{V}_T(\mathcal{F})$.

No algorithm can guarantee regret better than $\mathcal{V}_T(\mathcal{F})$.

Sequential Rademacher Complexity

Classical Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{x_1,\ldots,x_T \in \mathcal{X}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(x_t) \right]$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_T) \sim \mathrm{Unif}\left(\{\pm 1\}^T\right)$

Sequential Rademacher Complexity

Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{x_1,\ldots,x_T \in \mathcal{X}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(x_t) \right]$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_T) \sim \text{Unif}\left(\{\pm 1\}^T\right)$

Sequential Rademacher Complexity

Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Sequential Rademacher Complexity

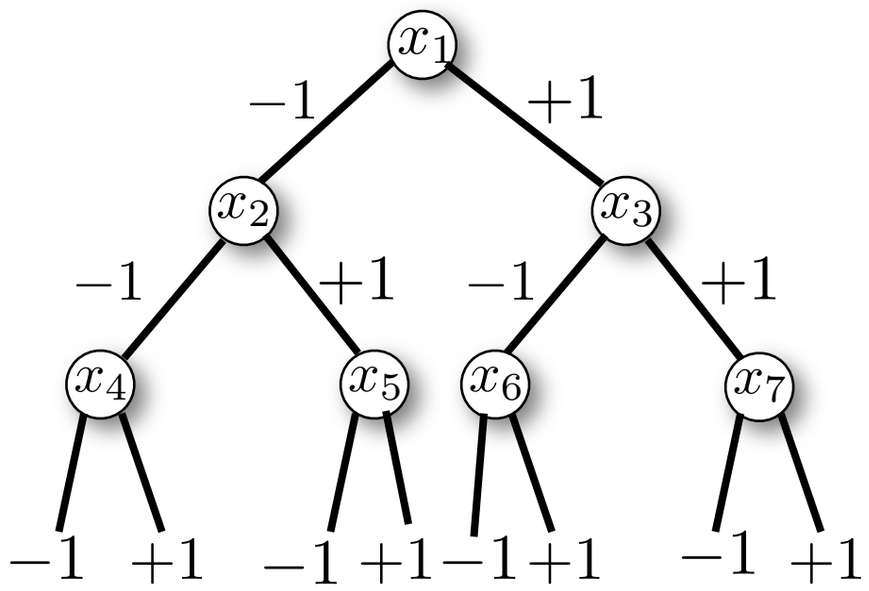
Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Definition : (tree)

An $\mathcal{X}$-valued tree of depth T is a sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of T mappings $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$



Sequential Rademacher Complexity

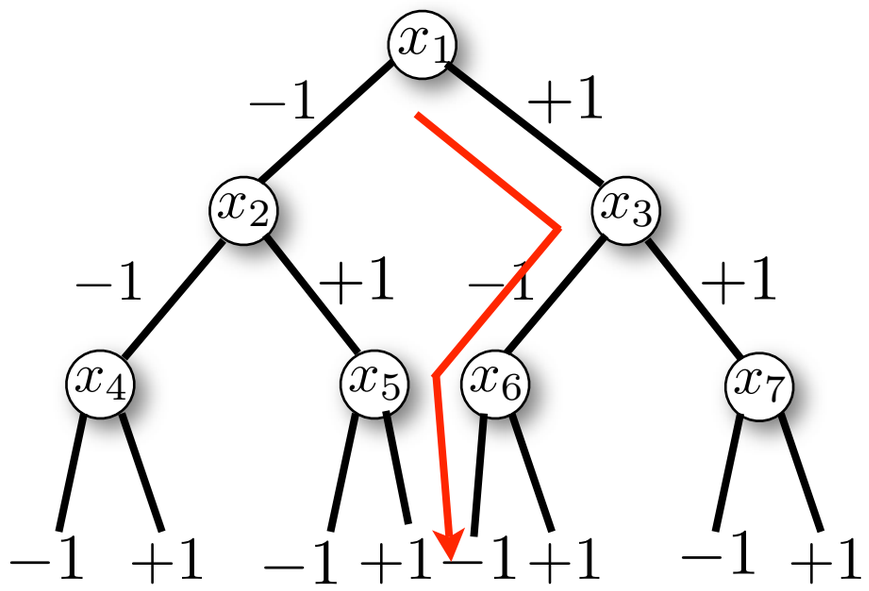
Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Definition : (tree)

An $\mathcal{X}$-valued tree of depth T is a sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of T mappings $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$



Example : $\epsilon = (+1, -1, -1)$

Sequential Rademacher Complexity

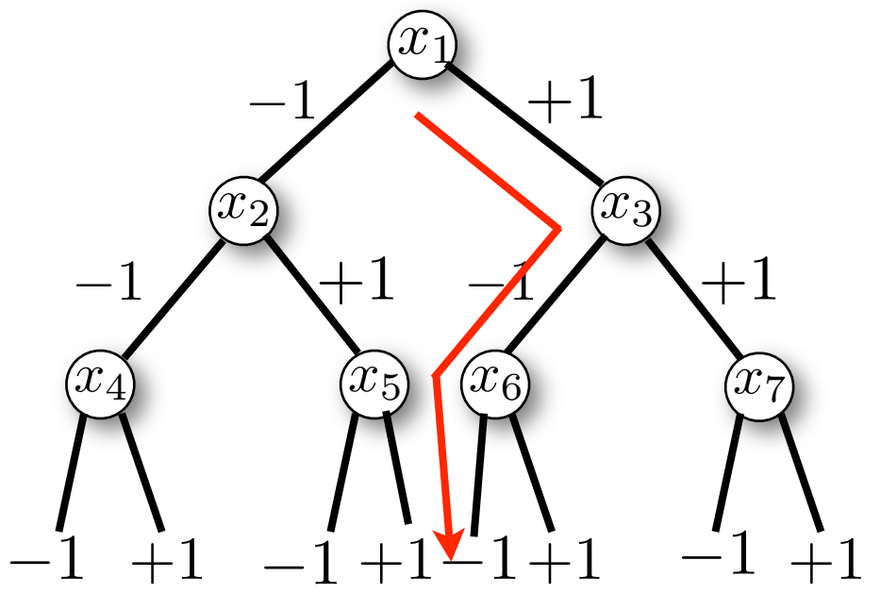
Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Definition : (tree)

An $\mathcal{X}$-valued tree of depth T is a sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of T mappings $\mathbf{x}_t : \{\pm 1\}^{t-1} \mapsto \mathcal{X}$



Example : $\epsilon = (+1, -1, -1)$

$$\sum_{t=1}^{3} \epsilon_t f(\mathbf{x}_t(\epsilon)) = +f(x_1) - f(x_3) - f(x_6)$$

Sequential Rademacher Complexity

Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Main Result :

$$\mathcal{V}_T(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Sequential Rademacher Complexity

Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Main Result :

$$\mathcal{V}_T(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Example : $\mathcal{F} = \{\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\| \leq 1\},\ \mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$

$\mathcal{R}_T(\mathcal{F}) \leq 2\sqrt{T}$ (online SVM)

Sequential Rademacher Complexity

Sequential Rademacher complexity :

$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Main Result :

$$\mathcal{V}_T(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Lipschitz contraction lemma and other basic properties hold.

Sequential Rademacher Complexity

Sequential Rademacher complexity :

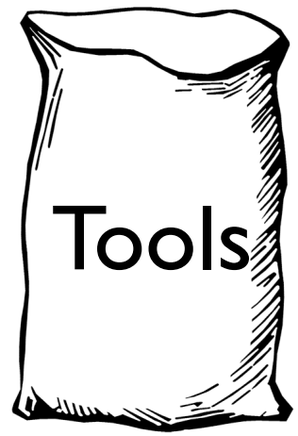
$$\mathcal{R}_T(\mathcal{F}) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

where supremum is over all $\mathcal{X}$-valued trees of depth T

Main Result :

$$\mathcal{V}_T(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Lipschitz contraction lemma and other basic properties hold.



Covering Numbers

Definition : (cover)

A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

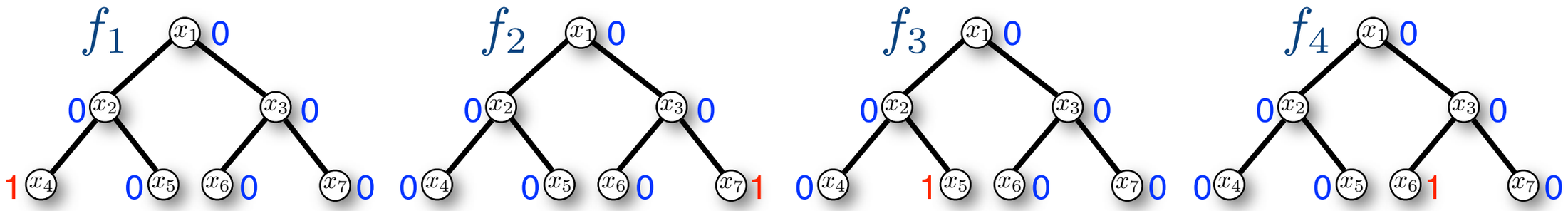
Covering Numbers

Definition : (cover)

A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

Example :



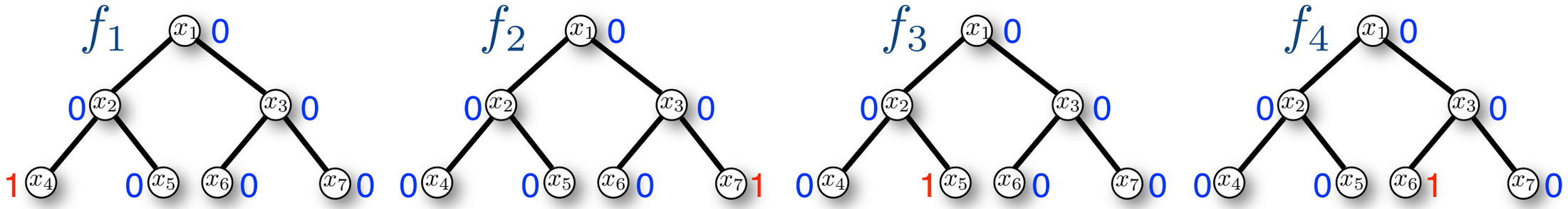
Covering Numbers

Definition : (cover)

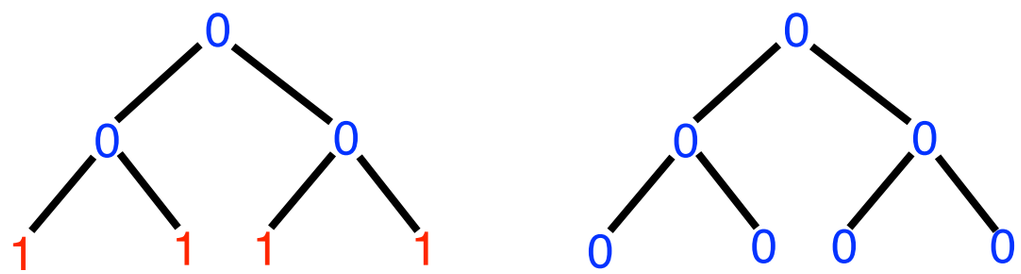
A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

Example :



covering trees :



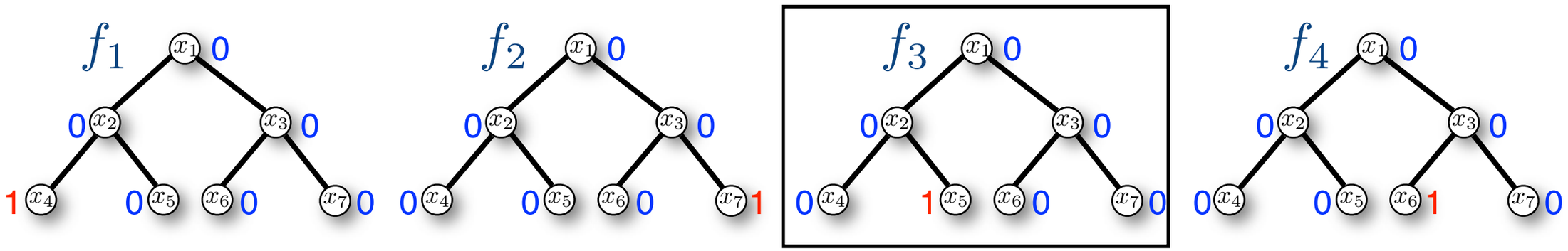
Covering Numbers

Definition : (cover)

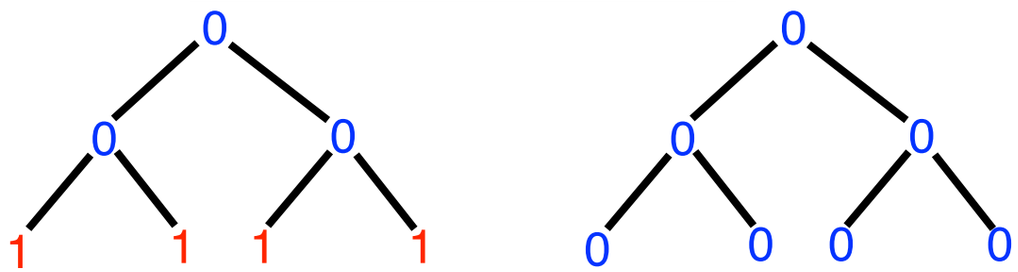
A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

Example :



covering trees :



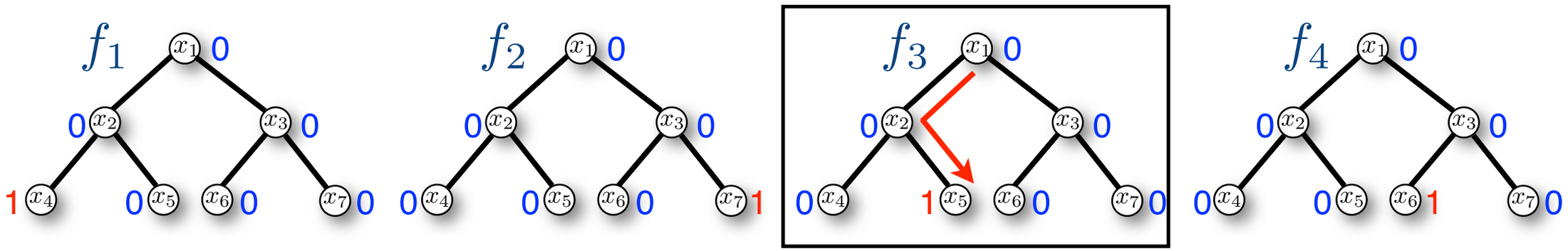
Covering Numbers

Definition : (cover)

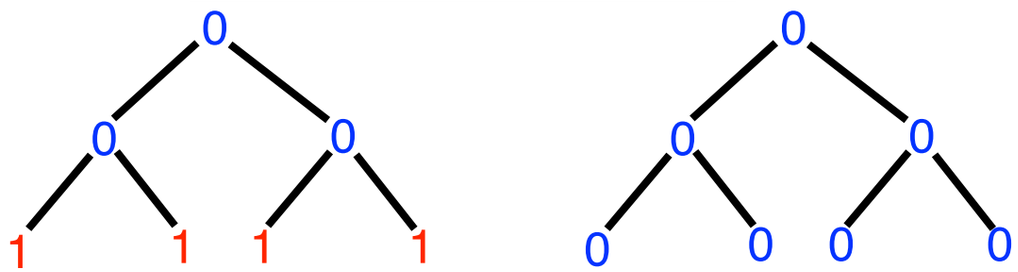
A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

Example :



covering trees :



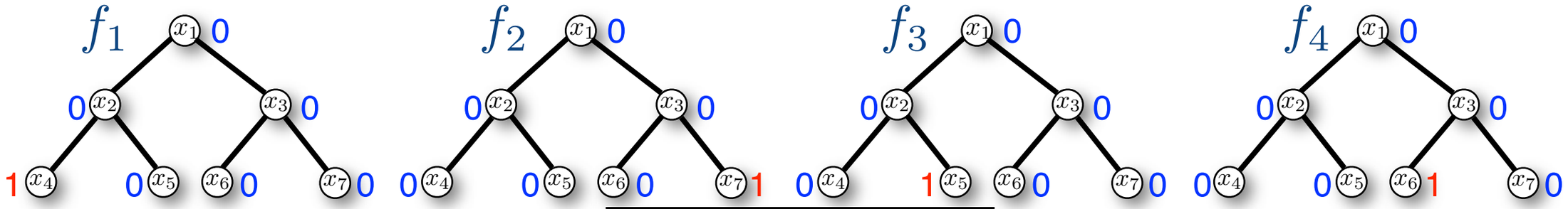
Covering Numbers

Definition : (cover)

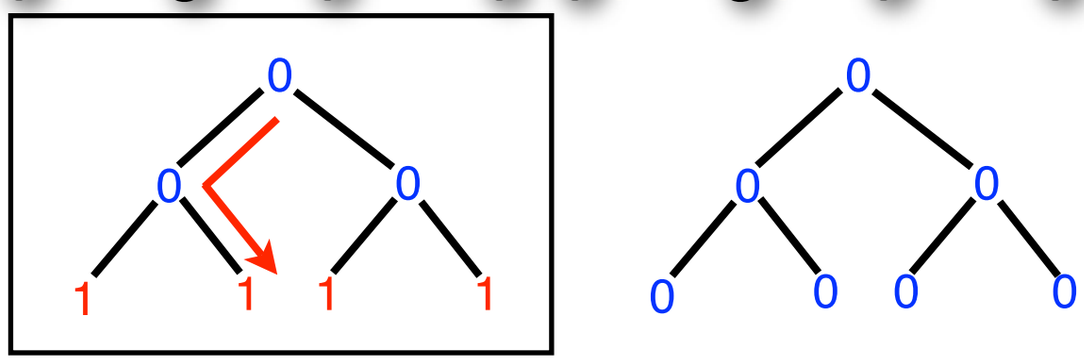
A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

Example :



covering trees :



Covering Numbers

Definition : (cover)

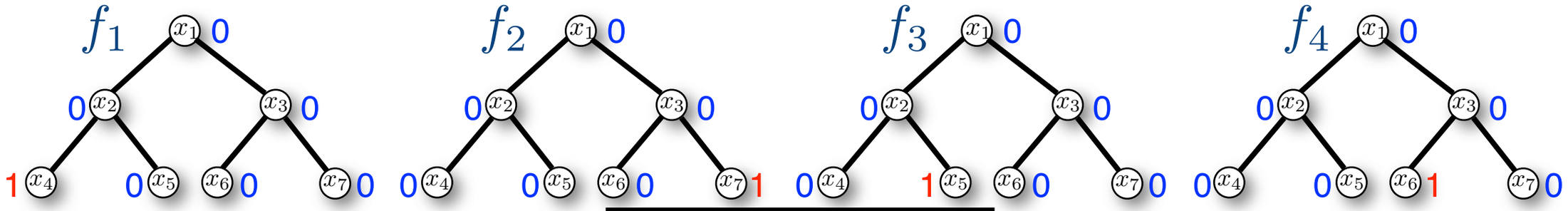
A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

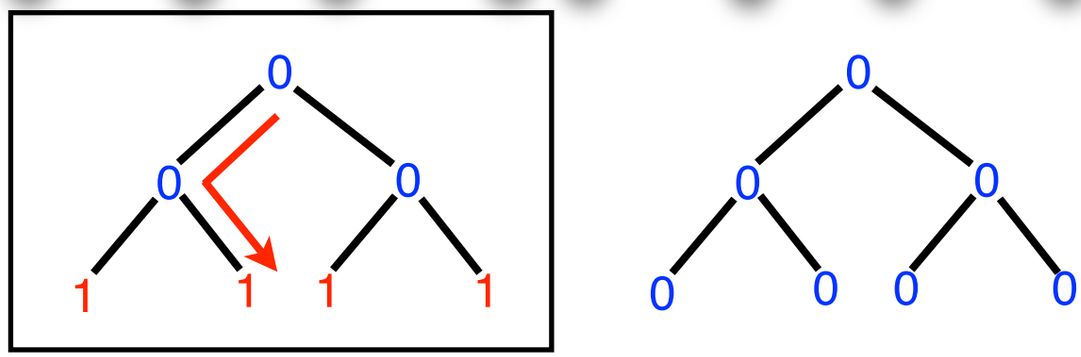
$\mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x}) =$ size of smallest cover V on tree $\mathbf{x}$

$$\mathcal{N}_p(\alpha, \mathcal{F}, T) = \sup_{\mathbf{x}} \mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x})$$

Example :



covering trees :



Covering Numbers

Definition : (cover)

A set V of $\mathbb{R}$-valued trees of depth T is an α cover (w.r.t. ℓ_p-norm) of $\mathcal{F}$ on a tree $\mathbf{x}$ of depth T if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^T,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{T} \sum_{t=1}^{T} |\mathbf{v}_t(\epsilon) - f(\mathbf{x}_t(\epsilon))|^p \right)^{1/p} \leq \alpha$$

$\mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x}) =$ size of smallest cover V on tree $\mathbf{x}$

$$\mathcal{N}_p(\alpha, \mathcal{F}, T) = \sup_{\mathbf{x}} \mathcal{N}_p(\alpha, \mathcal{F}, \mathbf{x})$$

How do we use covering numbers ?

Covering Numbers

Dudley integral complexity :

$$\mathcal{D}_T(\mathcal{F}) := \inf_{\alpha} \left\{ 8T\alpha + 24 \int_{\alpha}^{1} \sqrt{T \log \mathcal{N}_2(\delta, \mathcal{F}, T)} d\delta \right\}$$

Theorem : (Dudley integral bound)

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$,

$$\mathcal{V}_T(\mathcal{F}) \leq 2\mathcal{R}_T(\mathcal{F}) \leq \mathcal{D}_T(\mathcal{F})$$

Covering Numbers

Dudley integral complexity :

$$\mathcal{D}_T(\mathcal{F}) := \inf_{\alpha} \left\{ 8T\alpha + 24 \int_{\alpha}^{1} \sqrt{T \log \mathcal{N}_2(\delta, \mathcal{F}, T)} d\delta \right\}$$

Theorem : (Dudley integral bound)

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$,

$$\mathcal{V}_T(\mathcal{F}) \leq 2\mathcal{R}_T(\mathcal{F}) \leq \mathcal{D}_T(\mathcal{F})$$

Example : finite $\mathcal{F}$

$$\mathcal{D}_T(\mathcal{F}) \leq 24\sqrt{T \log |\mathcal{F}|}$$

Covering Numbers

Dudley integral complexity :

$$\mathcal{D}_T(\mathcal{F}) := \inf_{\alpha} \left\{ 8T\alpha + 24 \int_{\alpha}^{1} \sqrt{T \log \mathcal{N}_2(\delta, \mathcal{F}, T)} d\delta \right\}$$

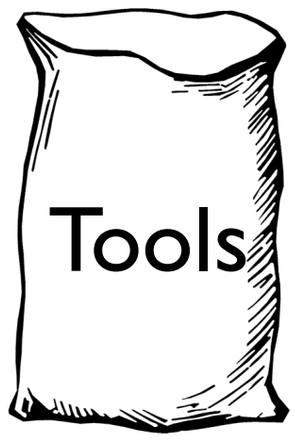
Theorem : (Dudley integral bound)

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$,

$$\mathcal{V}_T(\mathcal{F}) \leq 2\mathcal{R}_T(\mathcal{F}) \leq \mathcal{D}_T(\mathcal{F})$$

Example : finite $\mathcal{F}$

$$\mathcal{D}_T(\mathcal{F}) \leq 24\sqrt{T \log |\mathcal{F}|}$$



Combinatorial Parameters : Binary

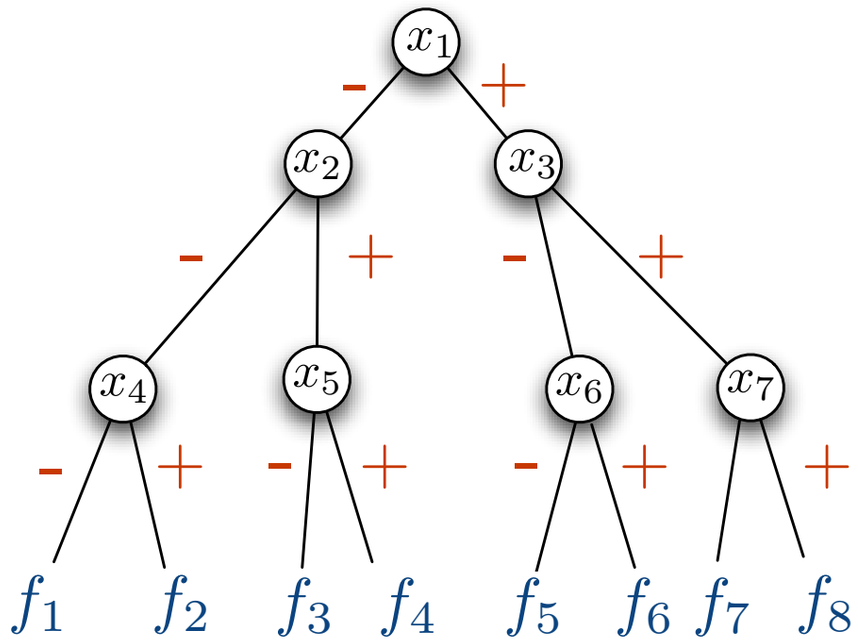
Definition : [Littlestone'88, Ben-David, Pal, Shalev-Shwartz'09]

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is shattered by $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ s.t. for all $t \in [d]$, $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$

Combinatorial Parameters : Binary

Definition : [Littlestone'88, Ben-David, Pal, Shalev-Shwartz'09]

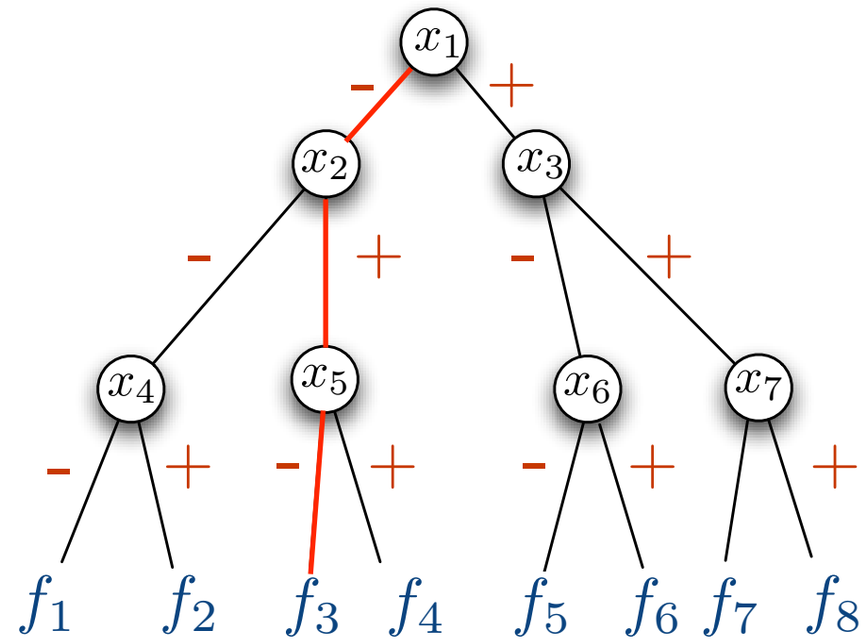
An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is shattered by $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ s.t. for all $t \in [d]$, $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$



Combinatorial Parameters : Binary

Definition : [Littlestone'88, Ben-David, Pal, Shalev-Shwartz'09]

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is shattered by $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ s.t. for all $t \in [d]$, $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$

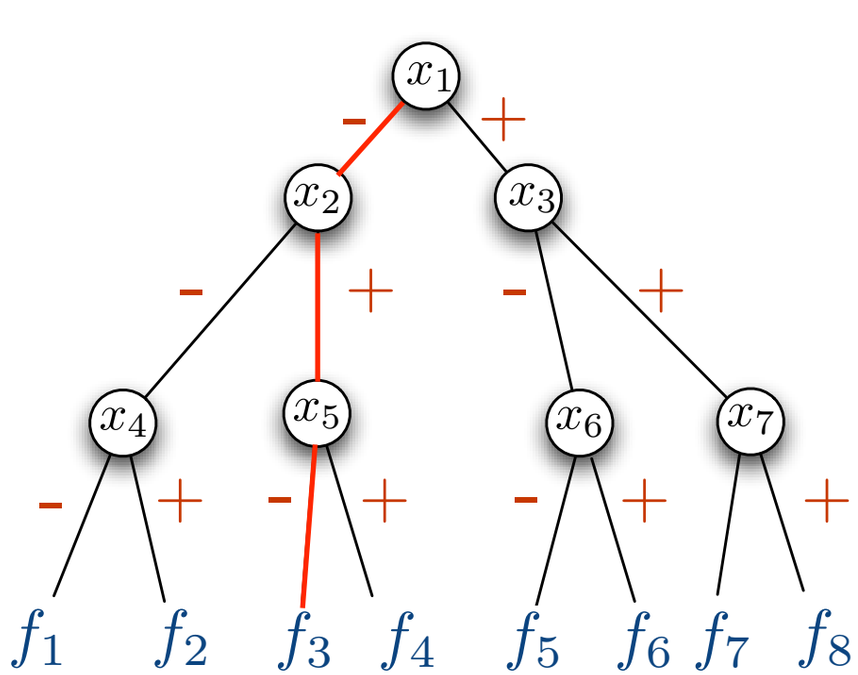


$$f_3(x_1) = -1,\ f_3(x_2) = +1,\ f_3(x_5) = -1$$

Combinatorial Parameters : Binary

Definition : [Littlestone'88, Ben-David, Pal, Shalev-Shwartz'09]

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is shattered by $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ s.t. for all $t \in [d]$, $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$



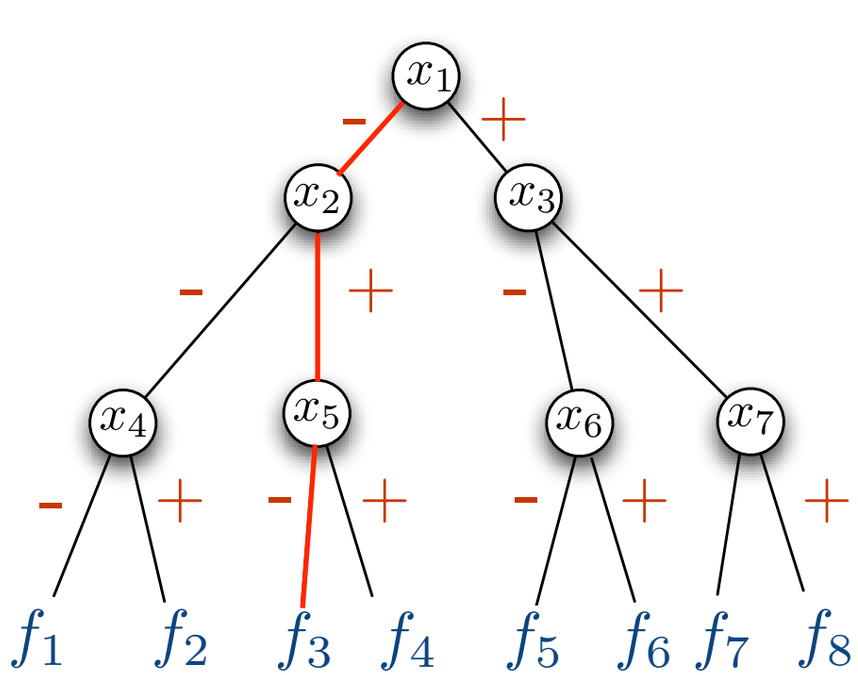
This tree is not the same object as covering tree!

Combinatorial Parameters : Binary

Definition : [Littlestone'88, Ben-David, Pal, Shalev-Shwartz'09]

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is shattered by $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ if for all $\epsilon \in \{\pm 1\}^d$, there exists $f \in \mathcal{F}$ s.t. for all $t \in [d]$, $f(\mathbf{x}_t(\epsilon)) = \epsilon_t$

Littlestone dimension $\mathrm{Ldim}(\mathcal{F})$ is the largest d s.t. $\mathcal{F}$ shatters some $\mathcal{X}$-valued tree of depth d



This tree is not the same object as covering tree!

Combinatorial Parameters

Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.

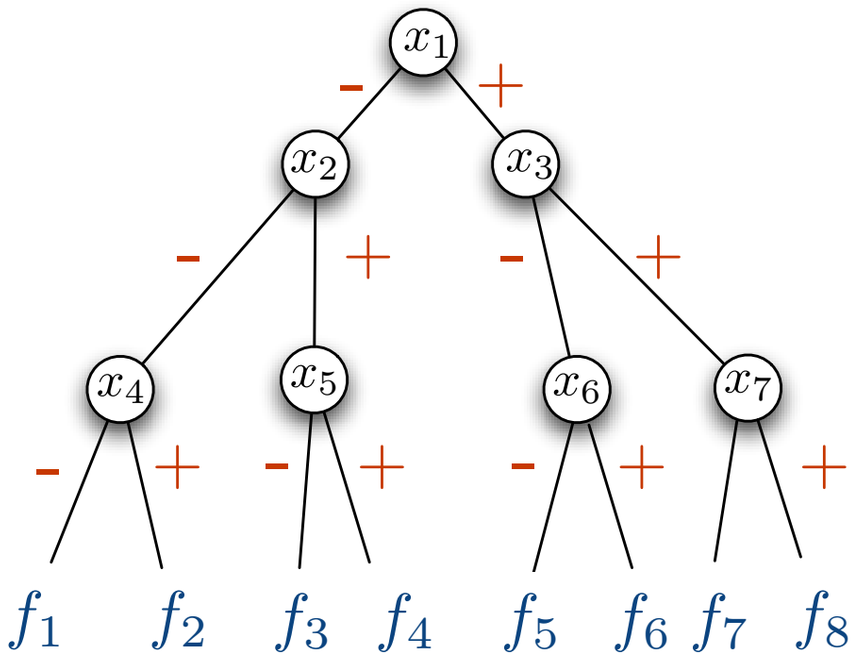
Combinatorial Parameters

Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.



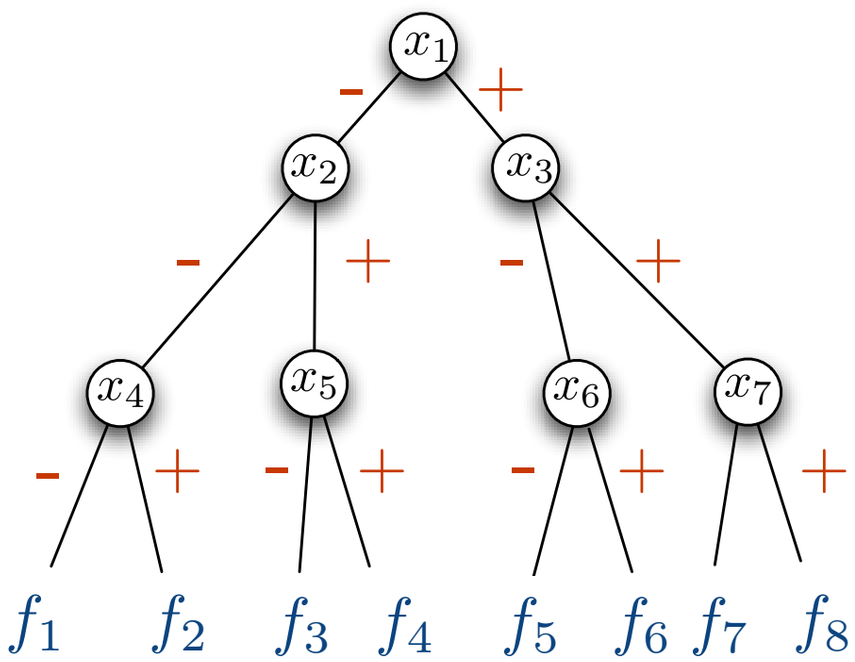
Combinatorial Parameters

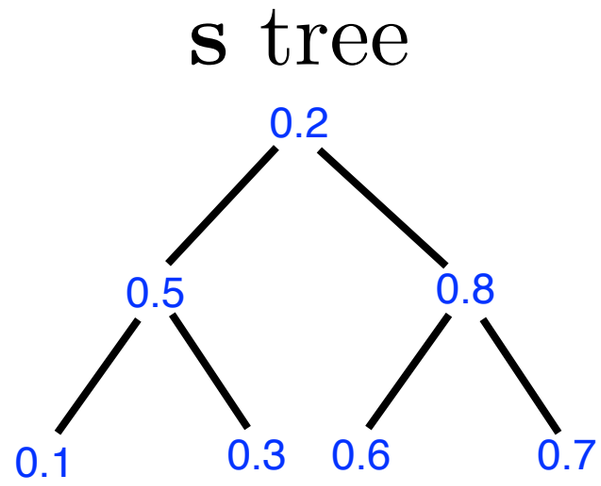
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





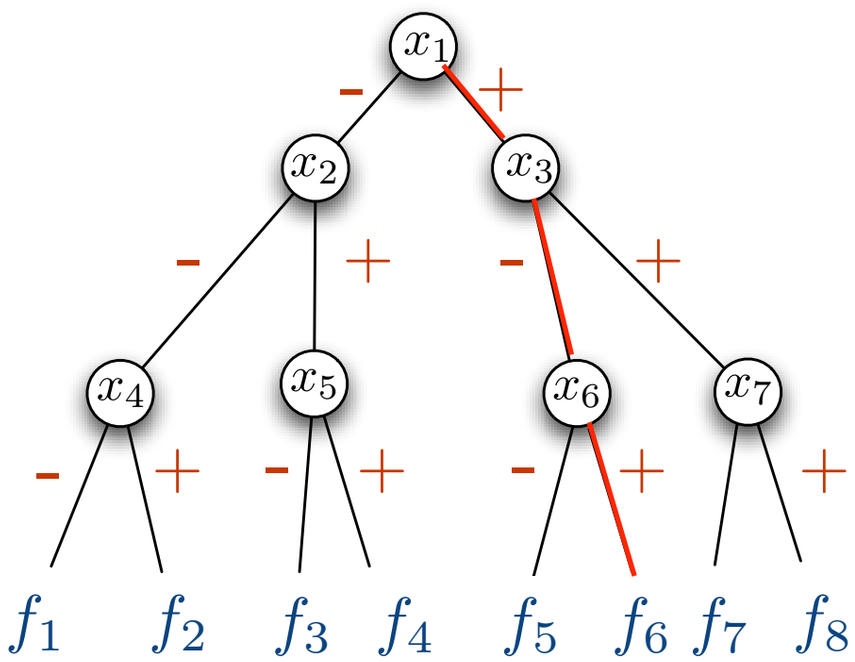
Combinatorial Parameters

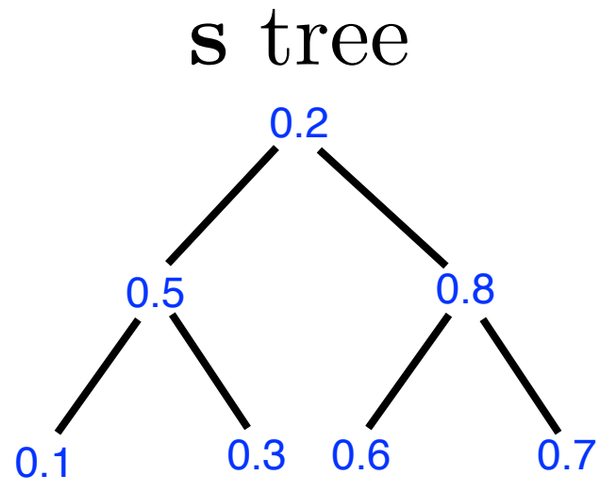
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





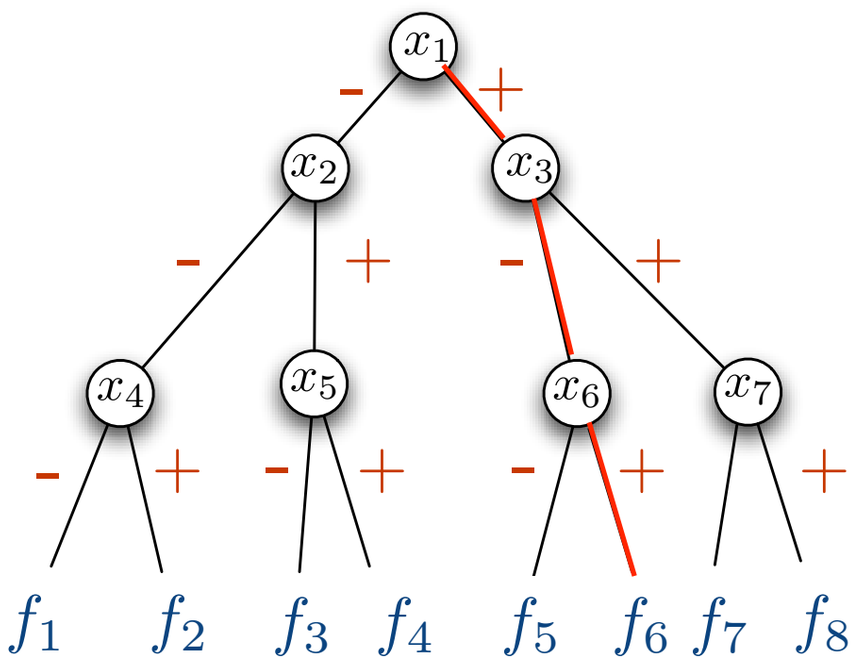
Combinatorial Parameters

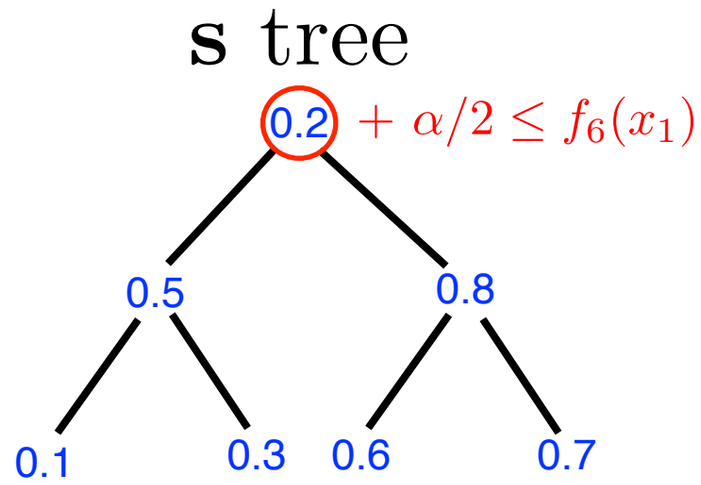
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





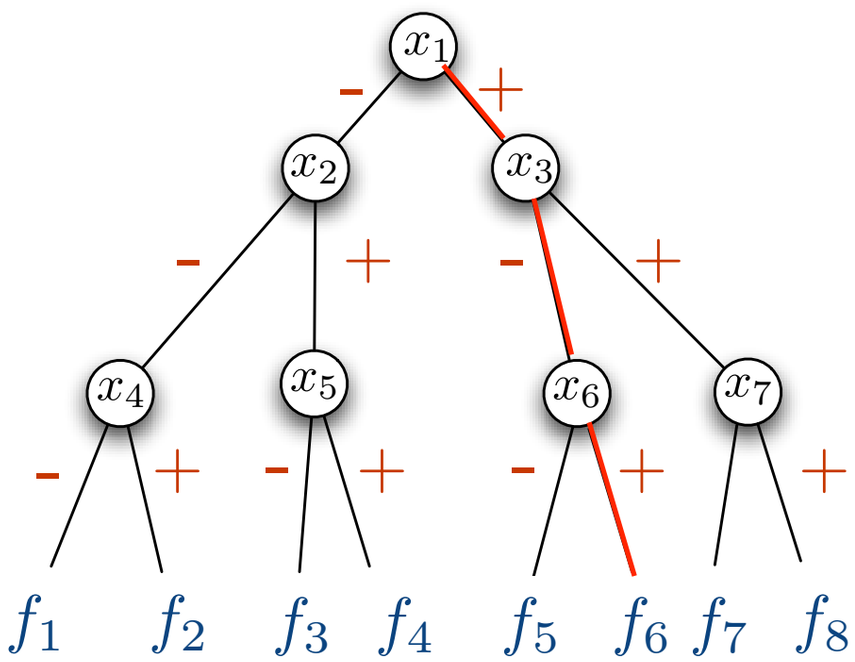
Combinatorial Parameters

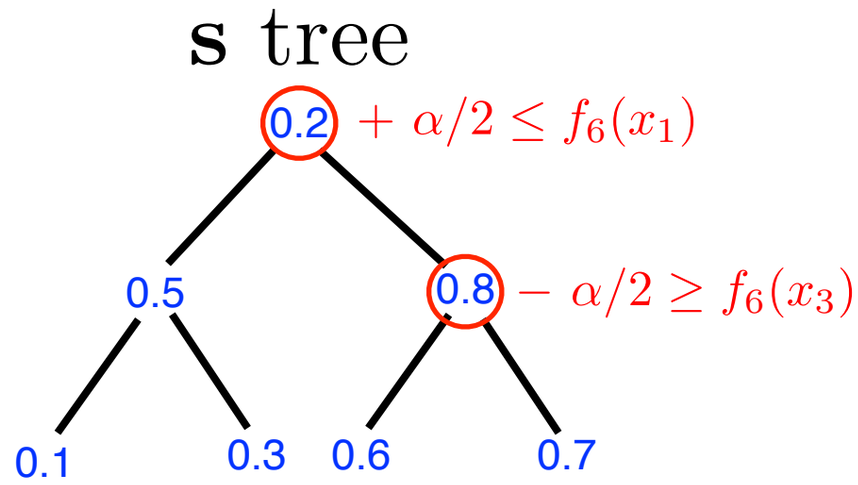
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





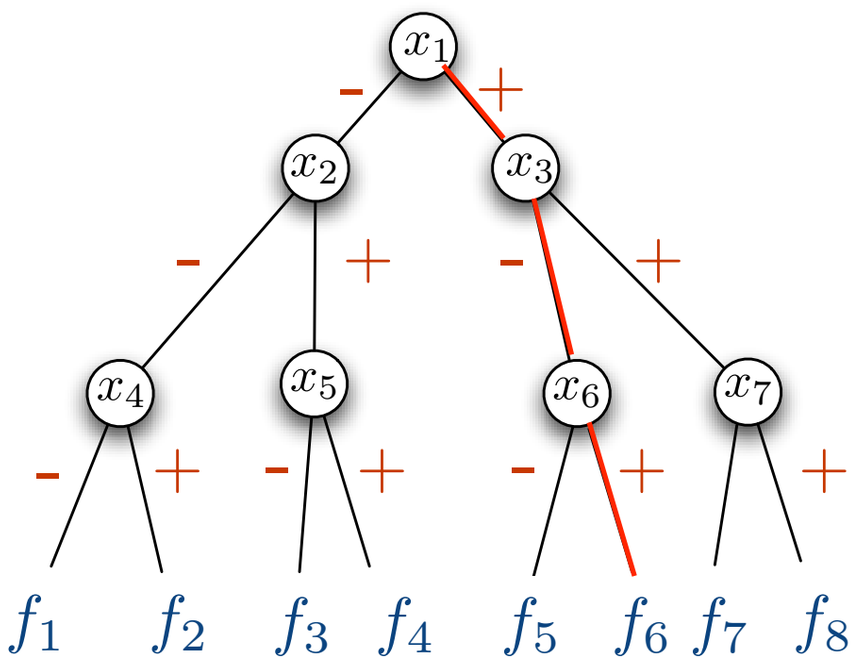
Combinatorial Parameters

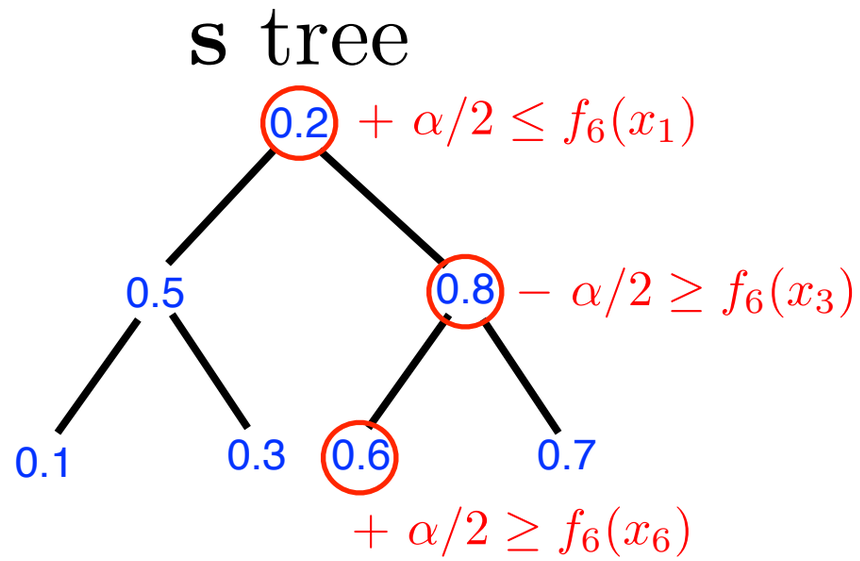
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





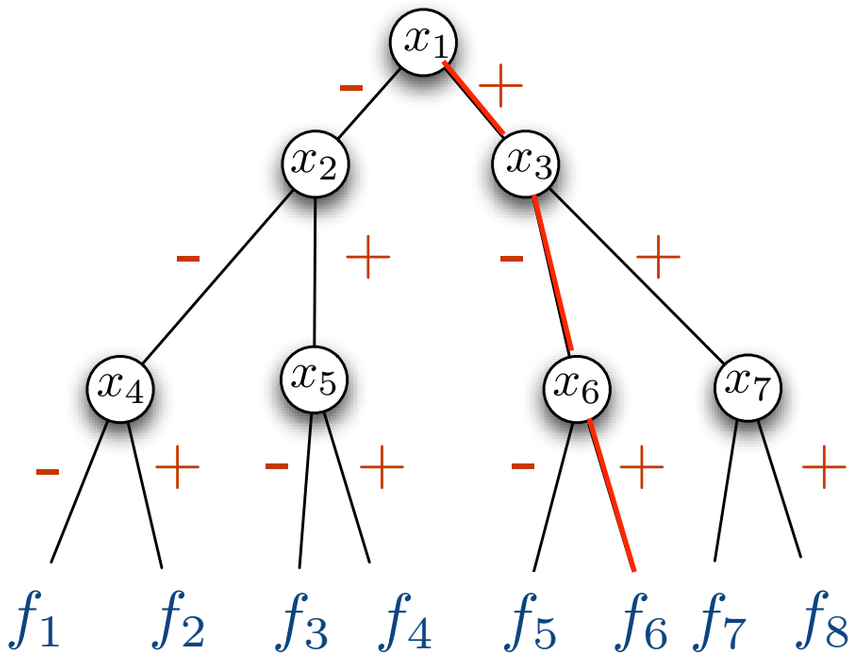
Combinatorial Parameters

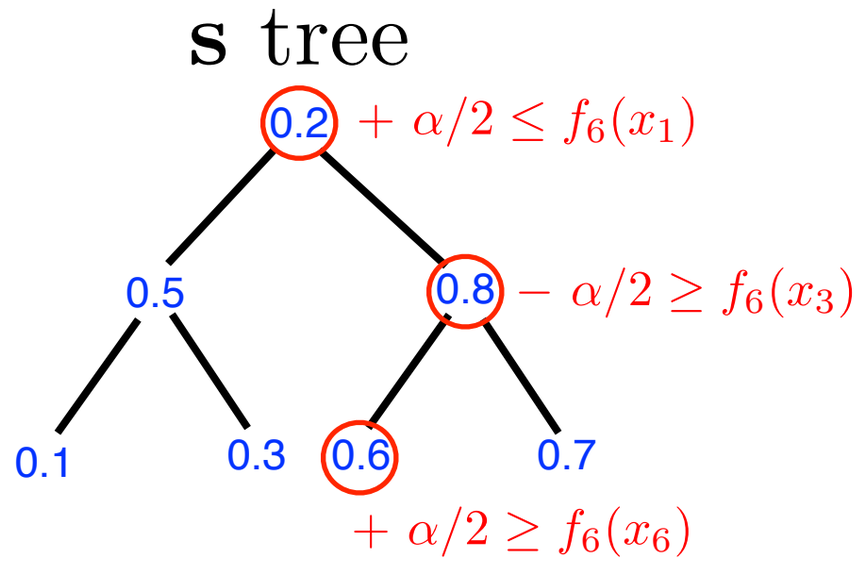
Definition : (fat-shattering)

An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth d is α-shattered by $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ if there exists a $\mathbb{R}$-valued tree $\mathbf{s}$ of depth d s.t.

$$\forall \epsilon \in \{\pm 1\}^d,\ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d],\quad \epsilon_t \left(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)\right) \geq \alpha/2$$

Fat-shattering dimension $\mathrm{fat}_\alpha(\mathcal{F})$ at scale α is the largest d s.t. $\mathcal{F}$ α-shatters some $\mathcal{X}$-valued tree of depth d.





How do we use this parameter ?

Analog to Sauer-Shelah Lemma

Theorem :

For any $\mathcal{F} \subset \{0, \ldots, k\}^{\mathcal{X}}$ with $\mathrm{fat}_1(\mathcal{F}) = d$:

$$\mathcal{N}(0, \mathcal{F}, T) \leq \sum_{i=0}^{d} \binom{T}{i} k^i \leq \left(\frac{ekT}{d}\right)^d$$

Analog to Sauer-Shelah Lemma

Theorem :

For any $\mathcal{F} \subset \{0, \ldots, k\}^{\mathcal{X}}$ with $\mathrm{fat}_1(\mathcal{F}) = d$:

$$\mathcal{N}(0, \mathcal{F}, T) \leq \sum_{i=0}^{d} \binom{T}{i} k^i \leq \left(\frac{ekT}{d}\right)^d$$

Analogous to result in statistical learning setting [Alon et al'97, Bartlett et al'96] :

Theorem :

For any $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ and $\alpha > 0$

$$\mathcal{N}_p(\alpha, \mathcal{F}, T) \leq \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})}$$

Analog to Sauer-Shelah Lemma

Theorem :

For any $\mathcal{F} \subset \{0, \ldots, k\}^{\mathcal{X}}$ with $\mathrm{fat}_1(\mathcal{F}) = d$:

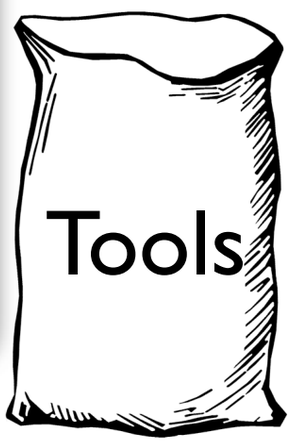
$$\mathcal{N}(0, \mathcal{F}, T) \leq \sum_{i=0}^{d} \binom{T}{i} k^i \leq \left(\frac{ekT}{d}\right)^d$$

Analogous to result in statistical learning setting [Alon et al'97, Bartlett et al'96] :

Theorem :

For any $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ and $\alpha > 0$

$$\mathcal{N}_p(\alpha, \mathcal{F}, T) \leq \left(\frac{2eT}{\alpha}\right)^{\mathrm{fat}_\alpha(\mathcal{F})}$$



Supervised Learning

Supervised Learning

Binary Classification :

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al.'89]

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al. '89]

Online learning : Learnable $\Leftrightarrow$ finite Ldim

[Ben-David, Pal, Shalev-Shwartz '09 ,Littlestone '88]

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al.'89]

Online learning : Learnable $\Leftrightarrow$ finite Ldim

[Ben-David, Pal, Shalev-Shwartz '09 ,Littlestone '88]

General Supervised Learning Problem : $\ell(f,(x,y)) = |f(x) - y|$

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al. '89]

Online learning : Learnable $\Leftrightarrow$ finite Ldim

[Ben-David, Pal, Shalev-Shwartz '09 ,Littlestone '88]

Can extend to other losses like squared loss ...

General Supervised Learning Problem : $\ell(f,(x,y)) = |f(x) - y|$

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al. '89]

Online learning : Learnable $\Leftrightarrow$ finite Ldim

[Ben-David, Pal, Shalev-Shwartz '09 ,Littlestone '88]

Can extend to other losses like squared loss ...

General Supervised Learning Problem : $\ell(f,(x,y)) = |f(x) - y|$

Statistical learning : Learnable $\Leftrightarrow$ $\forall \alpha > 0,\ \mathrm{fat}_\alpha < \infty$

[Alon et al '97, Bartlett et al. '96]

(classical)

Supervised Learning

Binary Classification :

Statistical learning : Learnable $\Leftrightarrow$ finite VCdim

[Vapnik, Chervonenkis '71, Blumer et al. '89]

Online learning : Learnable $\Leftrightarrow$ finite Ldim

[Ben-David, Pal, Shalev-Shwartz '09 ,Littlestone '88]

Can extend to other losses like squared loss ...

General Supervised Learning Problem : $\ell(f,(x,y)) = |f(x) - y|$

Statistical learning : Learnable $\Leftrightarrow$ $\forall \alpha > 0,\ \mathrm{fat}_\alpha < \infty$

[Alon et al '97, Bartlett et al. '96] (classical)

Analogous result for online supervised learning ?

Online Supervised Learning

Theorem :

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, following are equivalent

1. $\mathcal{F}$ is online learnable in the supervised setting.
2. For all $\alpha > 0$, $\quad \mathrm{fat}_\alpha < \infty$

Online Supervised Learning

Theorem :

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, following are equivalent

1. $\mathcal{F}$ is online learnable in the supervised setting.
2. For all $\alpha > 0$, $\quad \mathrm{fat}_\alpha < \infty$

Value of supervised game $\mathcal{V}_T^S(\mathcal{F})$, sequential Rademacher Complexity $\mathcal{R}_T(\mathcal{F})$ and Dudley-Integral complexity $\mathcal{D}_T(\mathcal{F})$ are all within a factor of $\mathcal{O}(\log^{3/2} T)$ of each other.

Online Supervised Learning

Theorem :

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, following are equivalent

1. $\mathcal{F}$ is online learnable in the supervised setting.
2. For all $\alpha > 0, \quad \text{fat}_\alpha < \infty$

Value of supervised game $\mathcal{V}_T^S(\mathcal{F})$, sequential Rademacher Complexity $\mathcal{R}_T(\mathcal{F})$ and Dudley-Integral complexity $\mathcal{D}_T(\mathcal{F})$ are all within a factor of $\mathcal{O}(\log^{3/2} T)$ of each other.

In fact : $$\mathcal{R}_T(\mathcal{F}) \leq \mathcal{V}_T^S(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Online Supervised Learning

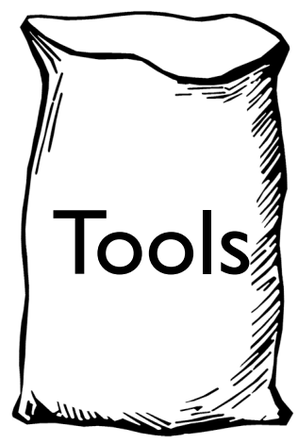
Theorem :

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, following are equivalent

1. $\mathcal{F}$ is online learnable in the supervised setting.
2. For all $\alpha > 0, \quad \text{fat}_\alpha < \infty$

Value of supervised game $\mathcal{V}_T^S(\mathcal{F})$, sequential Rademacher Complexity $\mathcal{R}_T(\mathcal{F})$ and Dudley-Integral complexity $\mathcal{D}_T(\mathcal{F})$ are all within a factor of $\mathcal{O}(\log^{3/2} T)$ of each other.

In fact : $\mathcal{R}_T(\mathcal{F}) \leq \mathcal{V}_T^S(\mathcal{F}) \leq 2\,\mathcal{R}_T(\mathcal{F})$



Online Supervised Learning

Theorem :

For any $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, following are equivalent

1. $\mathcal{F}$ is online learnable in the supervised setting.
2. For all $\alpha > 0, \quad \text{fat}_\alpha < \infty$

Value of supervised game $\mathcal{V}_T^S(\mathcal{F})$, sequential Rademacher Complexity $\mathcal{R}_T(\mathcal{F})$ and Dudley-Integral complexity $\mathcal{D}_T(\mathcal{F})$ are all within a factor of $\mathcal{O}(\log^{3/2} T)$ of each other.

In fact : $$\mathcal{R}_T(\mathcal{F}) \leq \mathcal{V}_T^S(\mathcal{F}) \leq 2\, \mathcal{R}_T(\mathcal{F})$$

Extending [Ben-David, Pal, Shalev-Shwartz '09] we provide Generic Algorithm for supervised learning.

Applications

We provide bounds for (non-constructive):

- Online convex optimization / Linear function classes
- Multi-layer Neural Networks
- Decision Trees
- Generic Margin Bounds
- Online Isotonic Regression and Regression with Classes of Lipschitz transformtion
- Online Transductive Learning
- Prediction of Individual Sequences

and more . . .

Applications

We provide bounds for (non-constructive):

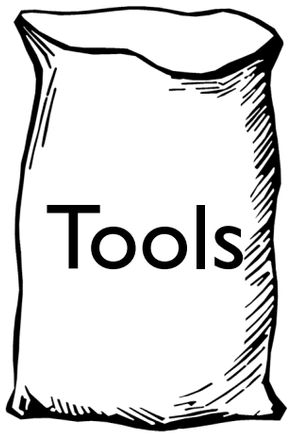
- Online convex optimization / Linear function classes
- Multi-layer Neural Networks
- Decision Trees
- Generic Margin Bounds
- Online Isotonic Regression and Regression with Classes of Lipschitz transformtion
- Online Transductive Learning
- Prediction of Individual Sequences

and more . . .

Come to poster T17!

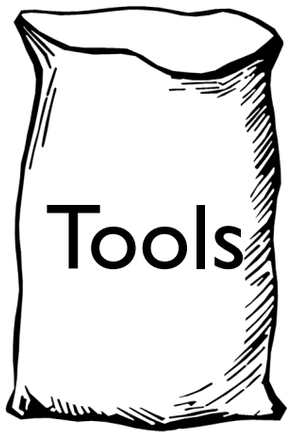
Summary

Summary



Summary

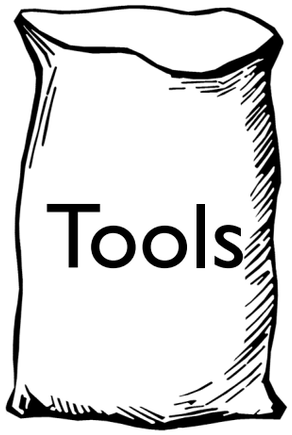
1. Sequential Rademacher complexity.



Summary

1. Sequential Rademacher complexity.

2. Dudley Integral complexity and tree based covering numbers.

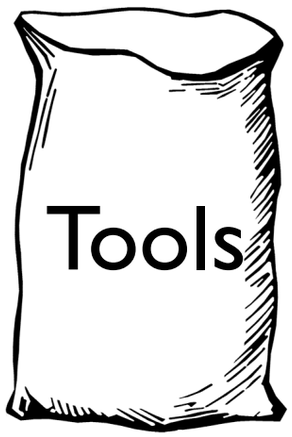


Summary

1. Sequential Rademacher complexity.

2. Dudley Integral complexity and tree based covering numbers.

3. Fat-shattering on trees dimension and Sauer-Shelah lemma.



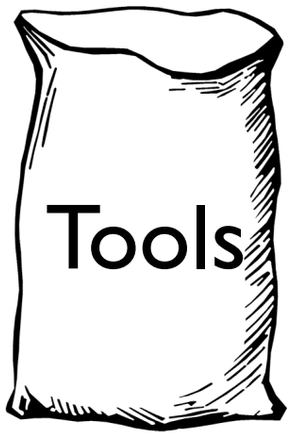
Summary

1. Sequential Rademacher complexity.

2. Dudley Integral complexity and tree based covering numbers.

3. Fat-shattering on trees dimension and Sauer-Shelah lemma.

4. Characterizing online supervised learning.



Discussion/Further Work

Discussion/Further Work

- Learning Vs Stochastic/constrained adversary

Discussion/Further Work

- Learning Vs Stochastic/constrained adversary
- Online Learning: Beyond Regret (arxiv version posted)

Discussion/Further Work

- Learning Vs Stochastic/constrained adversary
- Online Learning: Beyond Regret (arxiv version posted)
- Generic Algorithm whenever complexity is low?

Discussion/Further Work

- Learning Vs Stochastic/constrained adversary
- Online Learning: Beyond Regret (arxiv version posted)
- Generic Algorithm whenever complexity is low?
- Fast rate results?

Discussion/Further Work

- Learning Vs Stochastic/constrained adversary
- Online Learning: Beyond Regret (arxiv version posted)
- Generic Algorithm whenever complexity is low?
- Fast rate results?
- Efficient algorithms for interesting applications

Thanks!