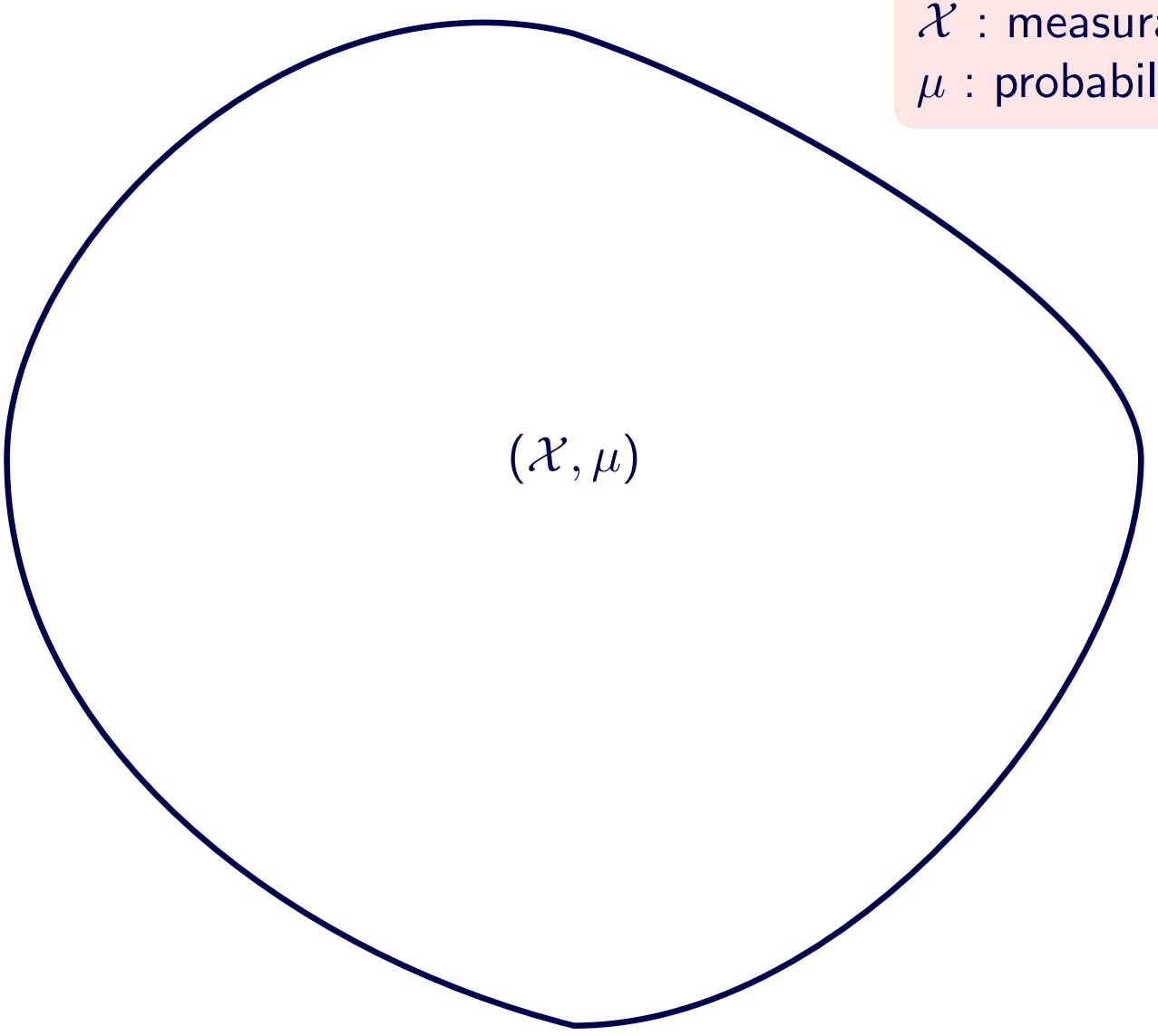
Aggregation and Empirical Risk Minimization

Guillaume Lecué
(joint works with Shahar Mendelson)

CNRS, LATP Marseille

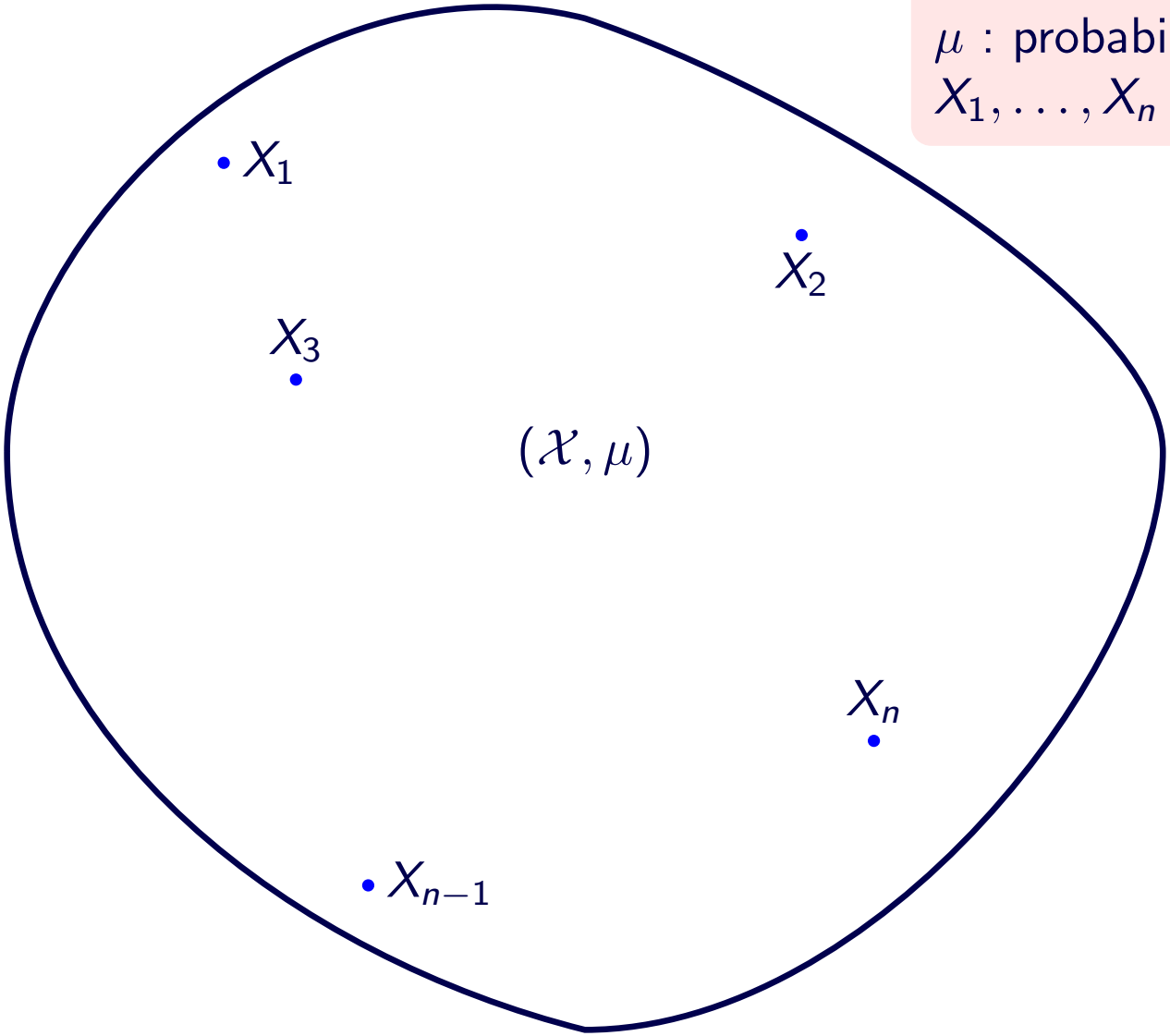
Model

$\mathcal{X}$: measurable space
μ : probability measure



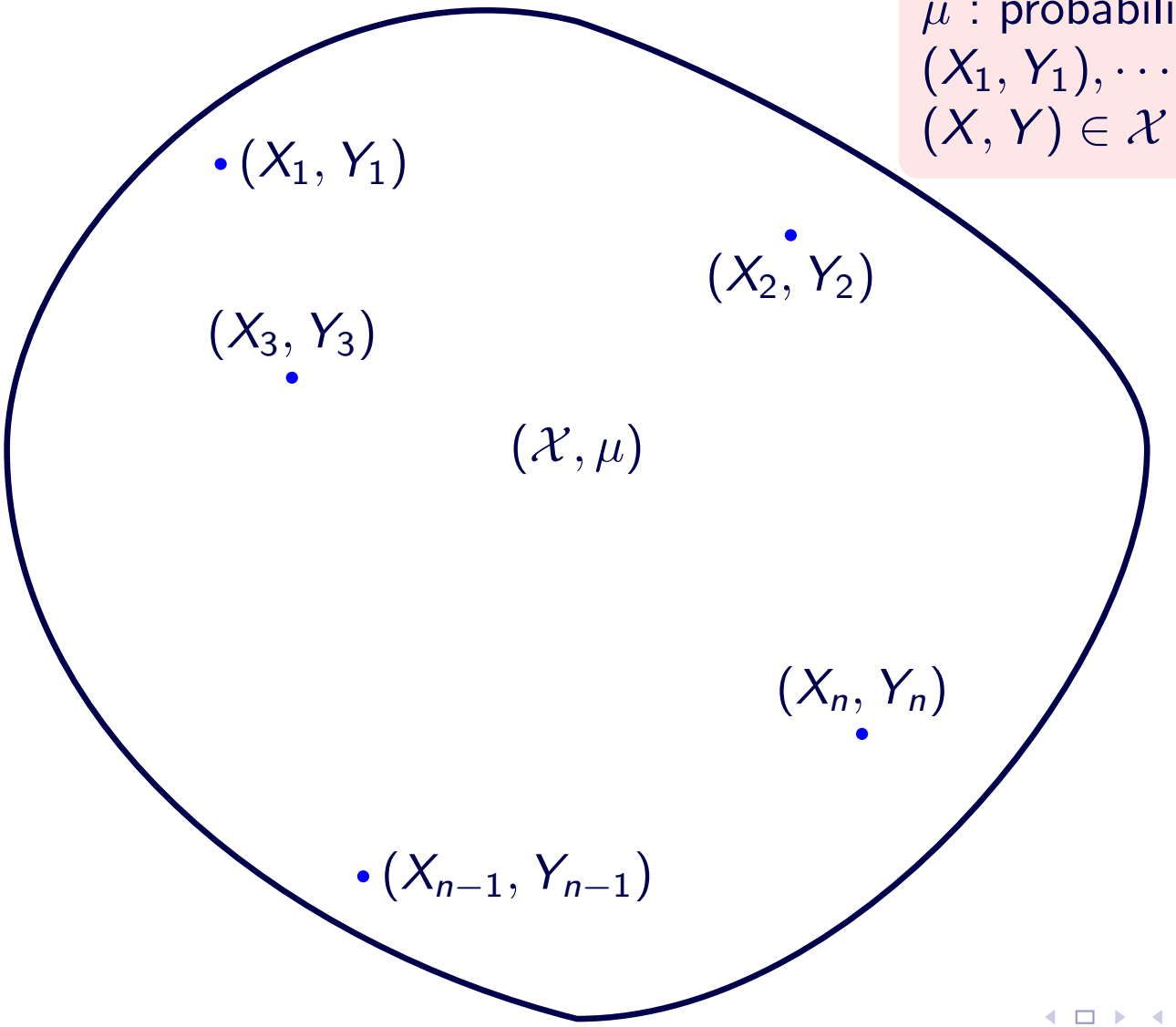
Model

$\mathcal{X}$: measurable space
μ : probability measure
$X_1, \ldots, X_n$ n i.i.d. r.v. $\sim \mu$



Model

$\mathcal{X}$: measurable space
μ : probability measure
$(X_1, Y_1), \cdots, (X_n, Y_n)$ i.i.d. r.v. $\sim$
$(X, Y) \in \mathcal{X} \times \mathbb{R} (\sim \nu)$



Problems

1. Problem of prediction

Problems

1. Problem of prediction

$$x \in \mathcal{X}$$

Problems

1. Problem of prediction

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

Problems

1. **Problem of prediction**

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.

Problems

1. **Problem of prediction**

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

Problems

1. **Problem of prediction**

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. **Agnostic learning**

Problems

1. **Problem of prediction**

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. **Agnostic learning**

No assumption on Y

F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

Problems

1. ## Problem of prediction

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. ## Agnostic learning

No assumption on Y
F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

$$\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$$

Problems

1. ## Problem of prediction

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. ## Agnostic learning

No assumption on Y
F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

$$\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$$

3. ## Problem of aggregation

Problems

1. ## Problem of prediction

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. ## Agnostic learning

No assumption on Y
F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

$$\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$$

3. ## Problem of aggregation

$|F| = M$;

Problems

1. **Problem of prediction**

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. **Agnostic learning**

No assumption on Y

F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

$$\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$$

3. **Problem of aggregation**

$|F| = M$; $F = \{f_1, \ldots, f_M\}$;

Problems

1. ## Problem of prediction

$$x \in \mathcal{X} \quad \longrightarrow \quad y \in \mathbb{R}$$

construction of $\hat{f}(\cdot, (X_i, Y_i)_i)$ s.t. $\mathbb{E}(\hat{f}(X) - Y)^2$ small.
$\mathbb{E}(\hat{f}(X) - Y)^2 = \mathbb{E}(\hat{f}(X) - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) - Y)^2$
$\Rightarrow \mathbb{E}(Y|\cdot) \in$ Class of functions (smooth, entropy,..)

2. ## Agnostic learning

No assumption on Y
F class of functions : $\mathcal{X} \longmapsto \mathbb{R} \Rightarrow$ best predictor of Y in F

$$\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$$

3. ## Problem of aggregation

$|F| = M$; $F = \{f_1, \ldots, f_M\}$; construction of a procedure which has a risk as close as possible to

$$\min_{j=1,\ldots,M} \mathbb{E}(f_j(X) - Y)^2$$

Examples of aggregation methods

Problem of aggregation : Construction of a procedure $\tilde{f}(\cdot, D_n)$ such that

$$\mathbb{E}(\hat{f}(X) - Y)^2 \leq \min_{f \in F} \mathbb{E}(f(X) - Y)^2 + r_n(F)$$

Examples of aggregation methods

Problem of aggregation : Construction of a procedure $\tilde{f}(\cdot, D_n)$ such that

$$\mathbb{E}(\hat{f}(X) - Y)^2 \leq \min_{f \in F} \mathbb{E}(f(X) - Y)^2 + r_n(F)$$

Empirical risk :

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

Examples of aggregation methods

Problem of aggregation : Construction of a procedure $\tilde{f}(\cdot, D_n)$ such that

$$\mathbb{E}(\hat{f}(X) - Y)^2 \leq \min_{f \in F} \mathbb{E}(f(X) - Y)^2 + r_n(F)$$

Empirical risk :

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

$$\mathbb{E}[R_n(f)] = \mathbb{E}(f(X) - Y)^2 := R(f)$$

Empirical risk minimization algorithm :

$$\hat{f}^{ERM} \in \operatorname{Arg}\min_{f \in F} R_n(f)$$

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

where $\operatorname{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

where $\operatorname{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} w_{j,n} f_j$$

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

where $\operatorname{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} w_{j,n} f_j \text{ where } w_{j,n} = \frac{\exp(-nR_n(f_j)/T)}{\sum_{k=1}^{M} \exp(-nR_n(f_k)/T)}$$

T : temperature parameter.

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

where $\operatorname{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} w_{j,n} f_j \text{ where } w_{j,n} = \frac{\exp(-nR_n(f_j)/T)}{\sum_{k=1}^{M} \exp(-nR_n(f_k)/T)}$$

T : temperature parameter.
Cumulative Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} \bar{w}_j f_j$$

Examples of aggregation methods

Empirical risk minimization algorithm over the convex hull of F :

$$\hat{f}^{ERMconv} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(F)} R_n(f),$$

where $\operatorname{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} w_{j,n} f_j \text{ where } w_{j,n} = \frac{\exp(-nR_n(f_j)/T)}{\sum_{k=1}^{M} \exp(-nR_n(f_k)/T)}$$

T : temperature parameter.
Cumulative Aggregation with exponential weights :

$$\tilde{f} := \sum_{j=1}^{M} \bar{w}_j f_j \text{ where } \bar{w}_j = \frac{1}{n} \sum_{p=1}^{n} w_{j,p}.$$

Optimal rate of aggregation

Definition

$\psi(n, M)$: optimal rate of aggregation with confidence $0 < \delta < 1/2$
$\tilde{f}$: optimal aggregation procedure with confidence δ

Optimal rate of aggregation

Definition

$\psi(n, M)$: optimal rate of aggregation with confidence $0 < \delta < 1/2$
$\tilde{f}$: optimal aggregation procedure with confidence δ

- $\forall n, M$, $\forall F$ of cardinality M and any target Y (all bounded by b), with ν^n-probability at least $1 - \delta$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(\delta)\psi(n, M),$$

Optimal rate of aggregation

Definition

$\psi(n, M)$: optimal rate of aggregation with confidence $0 < \delta < 1/2$
$\tilde{f}$: optimal aggregation procedure with confidence δ

- $\forall n, M$, $\forall F$ of cardinality M and any target Y (all bounded by b), with ν^n-probability at least $1 - \delta$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(\delta)\psi(n, M),$$

- $\exists c_2 > 0$ s.t. $\forall M, n$, $\forall \bar{f}$, $\exists F$ of cardinality M and a target Y (all bounded by b) such that, with ν^n-probability at least $1/2$,

$$R(\bar{f}) \geq \min_{f \in F} R(f) + c_2\psi(n, M).$$

Lower bound

$\exists c_2 > 0$ s.t. $\forall M, n, \forall \bar{f}, \exists F$ of cardinality M and a target Y (all bounded by b) such that, with ν^n-probability at least $1/2$,

$$R(\bar{f}) \geq \min_{f \in F} R(f) + c_2 \frac{\log M}{n}.$$

Lower bound

$\exists c_2 > 0$ s.t. $\forall M, n, \forall \bar{f}, \exists F$ of cardinality M and a target Y (all bounded by b) such that, with ν^n-probability at least $1/2$,

$$R(\bar{f}) \geq \min_{f \in F} R(f) + c_2 \frac{\log M}{n}.$$

Given n observations and a dictionary F of cardinality M, the minimum price that one has to pay to mimic the best function in F is at least

$$\frac{\log M}{n}$$

Lower bound

$\exists c_2 > 0$ s.t. $\forall M, n, \forall \bar{f}, \exists F$ of cardinality M and a target Y (all bounded by b) such that, with ν^n-probability at least $1/2$,

$$R(\bar{f}) \geq \min_{f \in F} R(f) + c_2 \frac{\log M}{n}.$$

Given n observations and a dictionary F of cardinality M, the minimum price that one has to pay to mimic the best function in F is at least

$$\frac{\log M}{n}$$

Is it possible to achieve this rate ?

Suboptimality of ERM over F and $\mathrm{Conv}(F)$

1. $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\hat{f}^{ERM}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

Suboptimality of ERM over F and $\mathrm{Conv}(F)$

1. $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\hat{f}^{ERM}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

2. $\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

Suboptimality of ERM over F and $\mathrm{Conv}(F)$

1. $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\hat{f}^{ERM}) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

2. $\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

$(1/\sqrt{n}) >> (\log M)/n \Longrightarrow \hat{f}^{ERM}$ and $\hat{f}^{ERMconv}$ cannot achieve the rate $(\log M)/n$.

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :

A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :

A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).

$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :

A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).

$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\Longrightarrow$ the ERM over F is optimal among all the selectors.

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :
A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).
$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\Longrightarrow$ the ERM over F is optimal among all the selectors.
There exists some convex combinations of the elements in F which are optimal aggregation procedures.

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :

A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).

$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\Longrightarrow$ the ERM over F is optimal among all the selectors.

There exists some convex combinations of the elements in F which are optimal aggregation procedures.

idea :

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :
A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).
$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\Longrightarrow$ the ERM over F is optimal among all the selectors.
There exists some convex combinations of the elements in F which are optimal aggregation procedures.
idea :
$\Longrightarrow$ the ERM over $\mathrm{Conv}(F)$ is optimal among all the convex aggregates ;

Suboptimality of ERM over $\mathrm{Conv}(F)$

First remark :
A Selector is an aggregation procedure taking its values in the dictionary F (for instance the ERM over F).
$\forall \bar{f}_n$(selector), $\forall n, M$, $\exists F$ of cardinality M such that, w.p.g. $1/2$

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + c\sqrt{\frac{\log M}{n}}$$

$\Longrightarrow$ the ERM over F is optimal among all the selectors.
There exists some convex combinations of the elements in F which are optimal aggregation procedures.
idea :
$\Longrightarrow$ the ERM over $\mathrm{Conv}(F)$ is optimal among all the convex aggregates ;
$\Longrightarrow$ the ERM over $\mathrm{Conv}(F)$ is an optimal aggregation procedure.

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
$\Longrightarrow$ the target of the ERM over $\mathrm{Conv}(F)$ is $\min_{f \in \mathrm{Conv}(F)} R(f)$.

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
$\Longrightarrow$ the target of the ERM over $\mathrm{Conv}(F)$ is $\min_{f \in \mathrm{Conv}(F)} R(f)$.

$$R(\hat{f}^{ERMconv}) \leq \min_{f \in \mathrm{Conv}(F)} R(f) + c\phi(M, n)$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
$\Longrightarrow$ the target of the ERM over $\mathrm{Conv}(F)$ is $\min_{f \in \mathrm{Conv}(F)} R(f)$.

$$\begin{aligned} R(\hat{f}^{ERMconv}) &\leq \min_{f \in \mathrm{Conv}(F)} R(f) + c\phi(M,n) \\ &\leq \min_{f \in F} R(f) + \left[c\phi(M,n) + \min_{f \in F} R(f) - \min_{f \in \mathrm{Conv}(F)} R(f)\right] \end{aligned}$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
$\Longrightarrow$ the target of the ERM over $\mathrm{Conv}(F)$ is $\min_{f \in \mathrm{Conv}(F)} R(f)$.

$$\begin{aligned} R(\hat{f}^{ERMconv}) &\leq \min_{f \in \mathrm{Conv}(F)} R(f) + c\phi(M,n) \\ &\leq \min_{f \in F} R(f) + \left[c\phi(M,n) + \min_{f \in F} R(f) - \min_{f \in \mathrm{Conv}(F)} R(f)\right] \end{aligned}$$

$\Longrightarrow$ gain in the approximation error : $\min_{f \in F} R(f) - \min_{f \in \mathrm{Conv}(F)} R(f)$;

Suboptimality of ERM over $\mathrm{Conv}(F)$

Second remark :

$$\hat{f}^{ERMconv} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(F)} R_n(f),$$

where $\mathrm{Conv}(F) = \{\sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geq 0 \text{ and } \sum \lambda_j = 1\}$.
$\Longrightarrow$ the target of the ERM over $\mathrm{Conv}(F)$ is $\min_{f \in \mathrm{Conv}(F)} R(f)$.

$$\begin{aligned} R(\hat{f}^{ERMconv}) &\leq \min_{f \in \mathrm{Conv}(F)} R(f) + c\phi(M,n) \\ &\leq \min_{f \in F} R(f) + \left[c\phi(M,n) + \min_{f \in F} R(f) - \min_{f \in \mathrm{Conv}(F)} R(f)\right] \end{aligned}$$

$\Longrightarrow$ gain in the approximation error : $\min_{f \in F} R(f) - \min_{f \in \mathrm{Conv}(F)} R(f)$;
$\Longrightarrow$ the ERM over $\mathrm{Conv}(F)$ is an optimal aggregation procedure.

Suboptimality of ERM over $\mathrm{Conv}(F)$

$\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

$\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

$(\phi_i)_{i \in \mathbb{N}}$ sequence of i.i.d Rademacher r.v.

$$F := \{0, \pm\phi_1, \ldots, \pm\phi_M\} \text{ and target } : Y := \phi_{M+1}$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

$\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

$(\phi_i)_{i \in \mathbb{N}}$ sequence of i.i.d Rademacher r.v.

$$F := \{0, \pm\phi_1, \ldots, \pm\phi_M\} \text{ and target } : Y := \phi_{M+1}$$

1. no gain in the approximation error

$$\min_{f \in F} R(f) = \min_{f \in \mathrm{Conv}(F)} R(f)$$

Suboptimality of ERM over $\mathrm{Conv}(F)$

$\forall n$, $\exists F$ of cardinality $c\sqrt{n}$ such that, w.p.g. $1/2$

$$R(\hat{f}^{ERMconv}) \geq \min_{f \in F} R(f) + \sqrt{\frac{c}{n}}.$$

$(\phi_i)_{i \in \mathbb{N}}$ sequence of i.i.d Rademacher r.v.

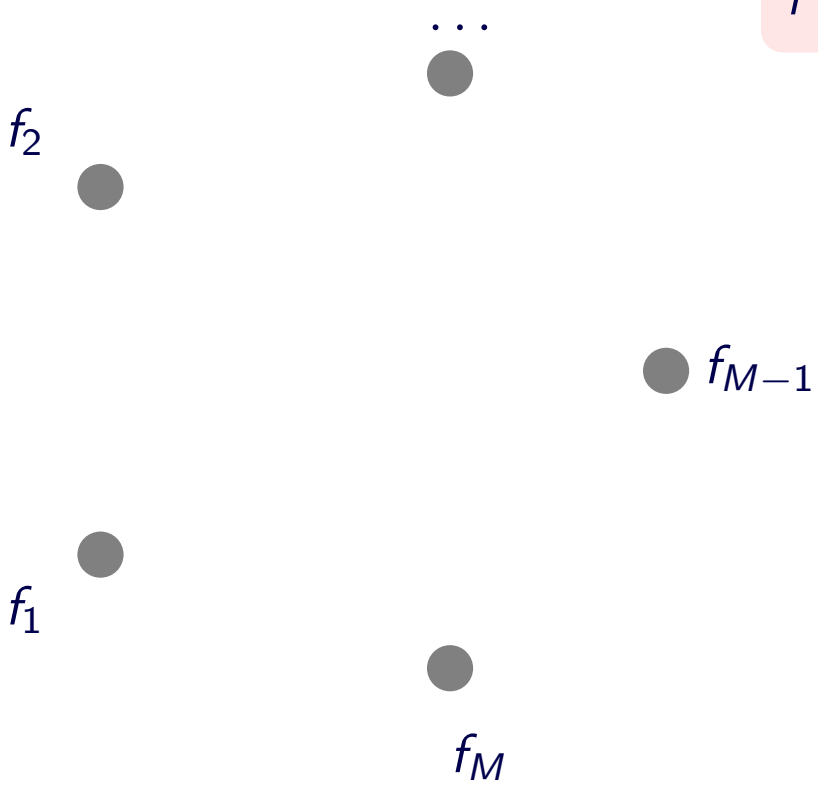
$$F := \{0, \pm\phi_1, \ldots, \pm\phi_M\} \text{ and target } : Y := \phi_{M+1}$$

1. no gain in the approximation error

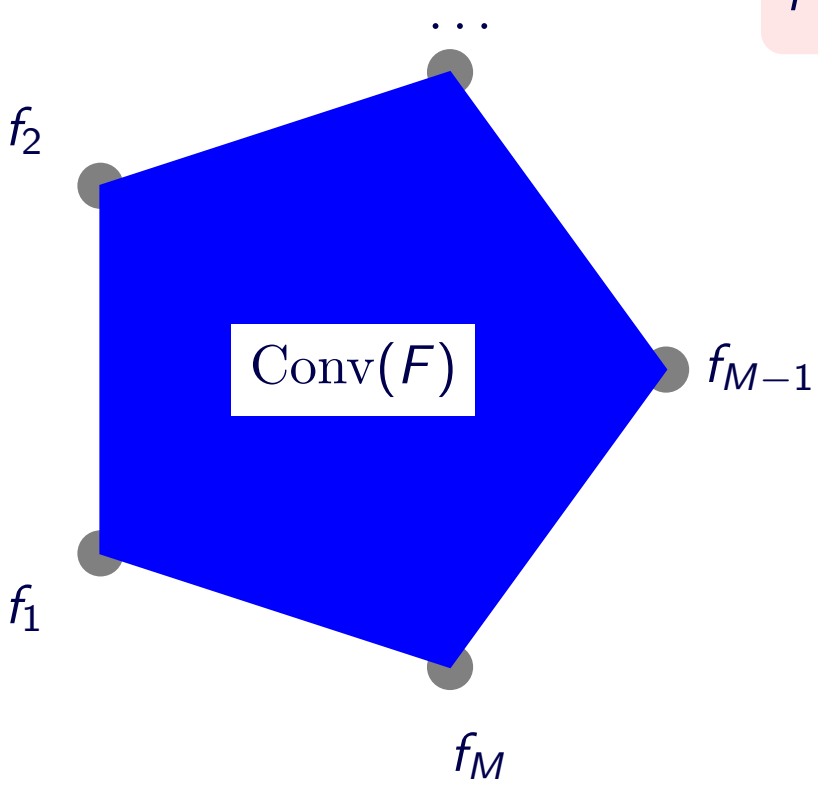
$$\min_{f \in F} R(f) = \min_{f \in \mathrm{Conv}(F)} R(f)$$

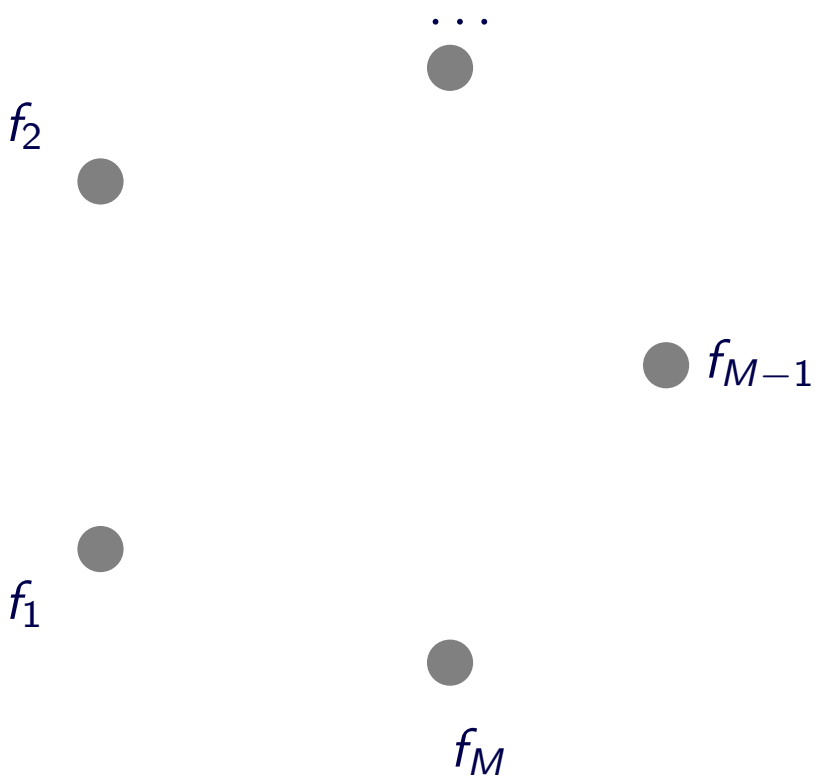
2. maximization of the complexity of $\mathrm{Conv}(F)$

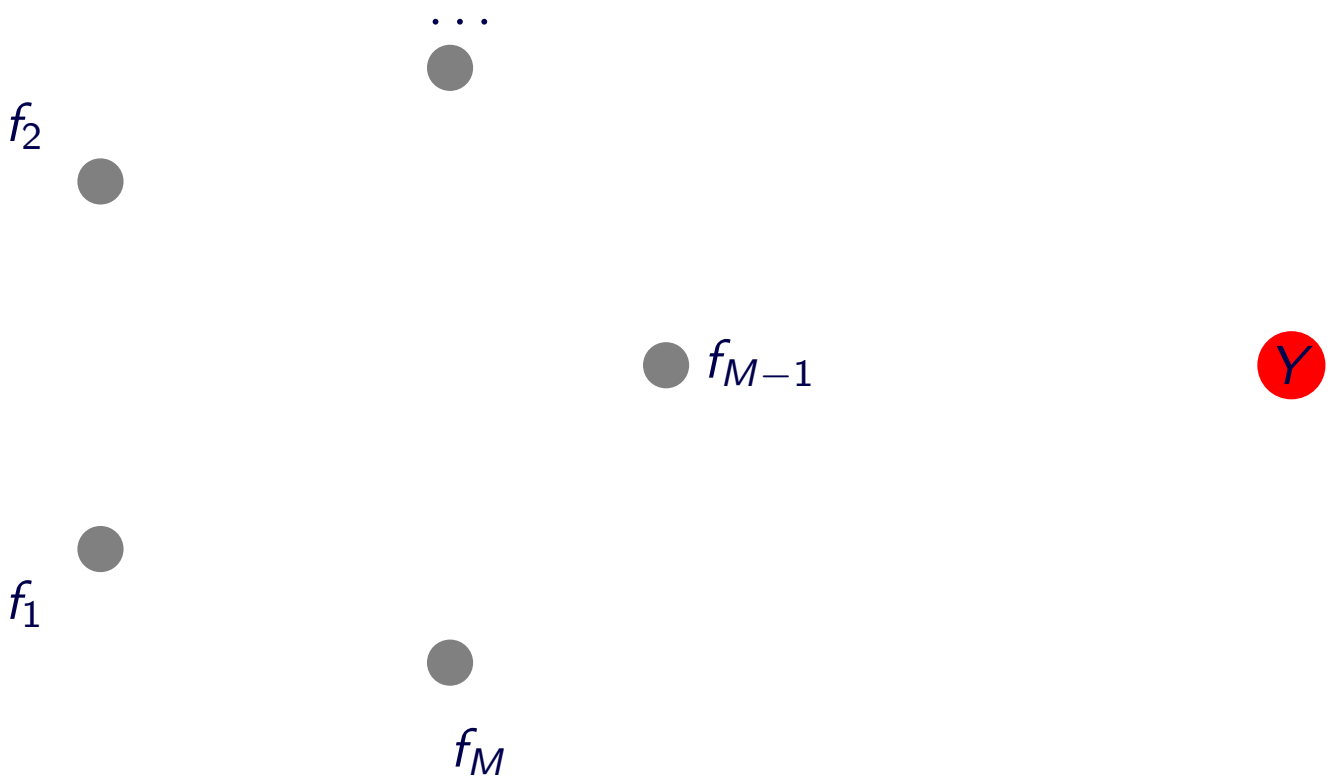
$\hat{f}^{ERM} \in \operatorname{Arg\,min}_{f \in F} R_n(f)$

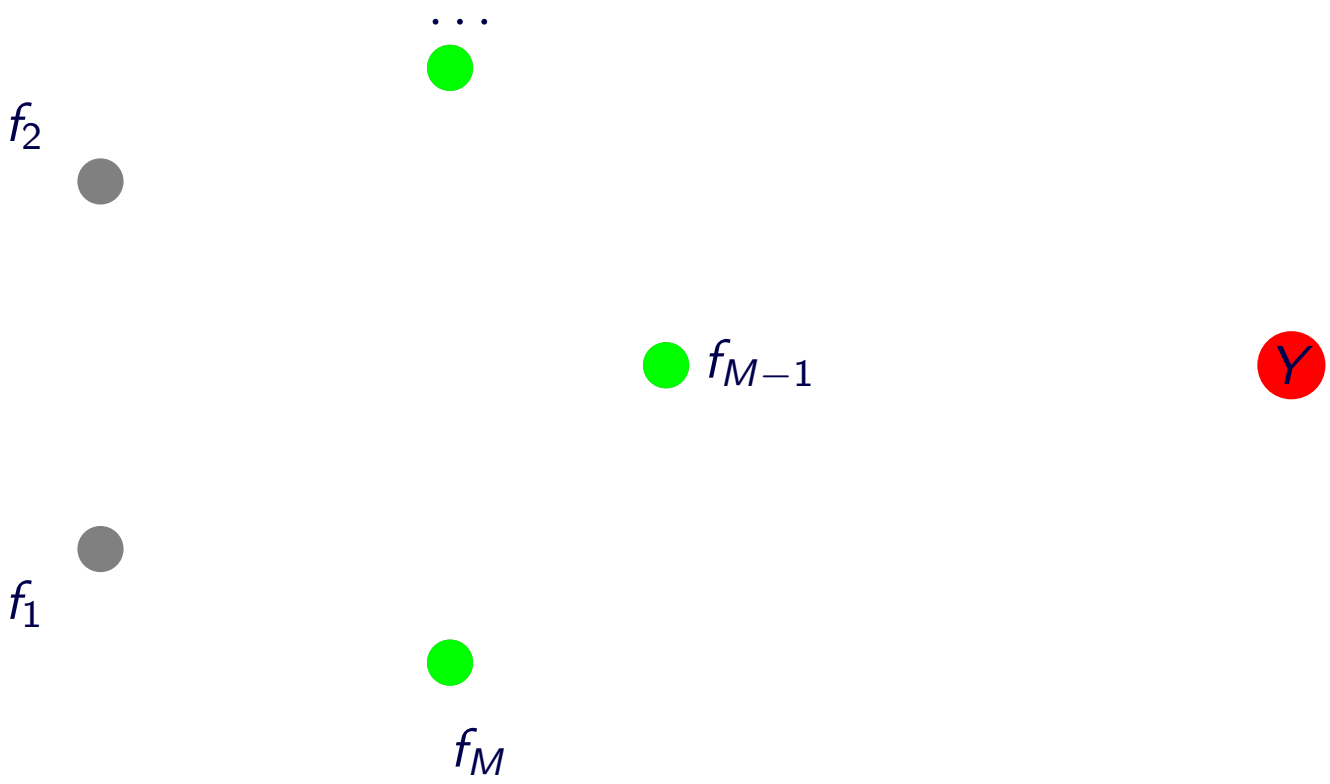


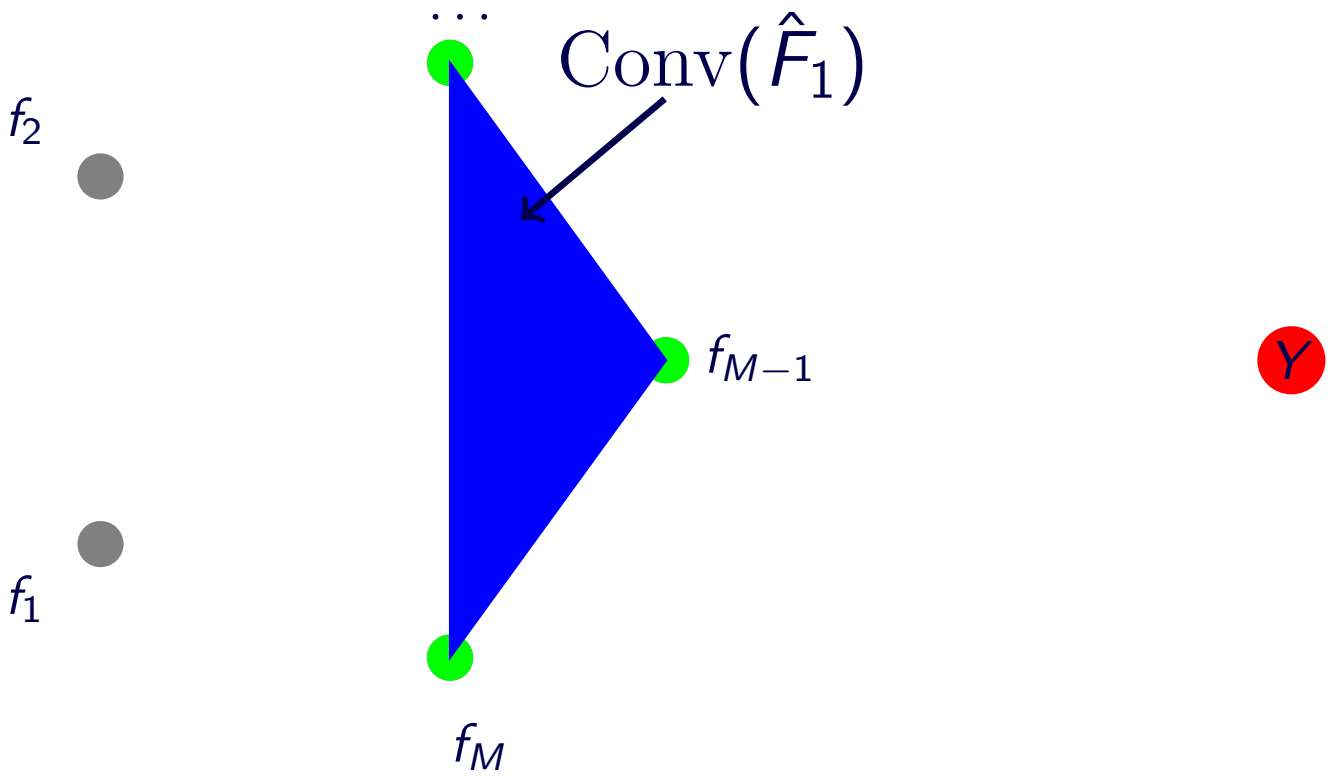
$$\hat{f}^{ERMconv} \in \operatorname{Arg\,min}_{f \in \mathrm{Conv}(F)} R_n(f)$$



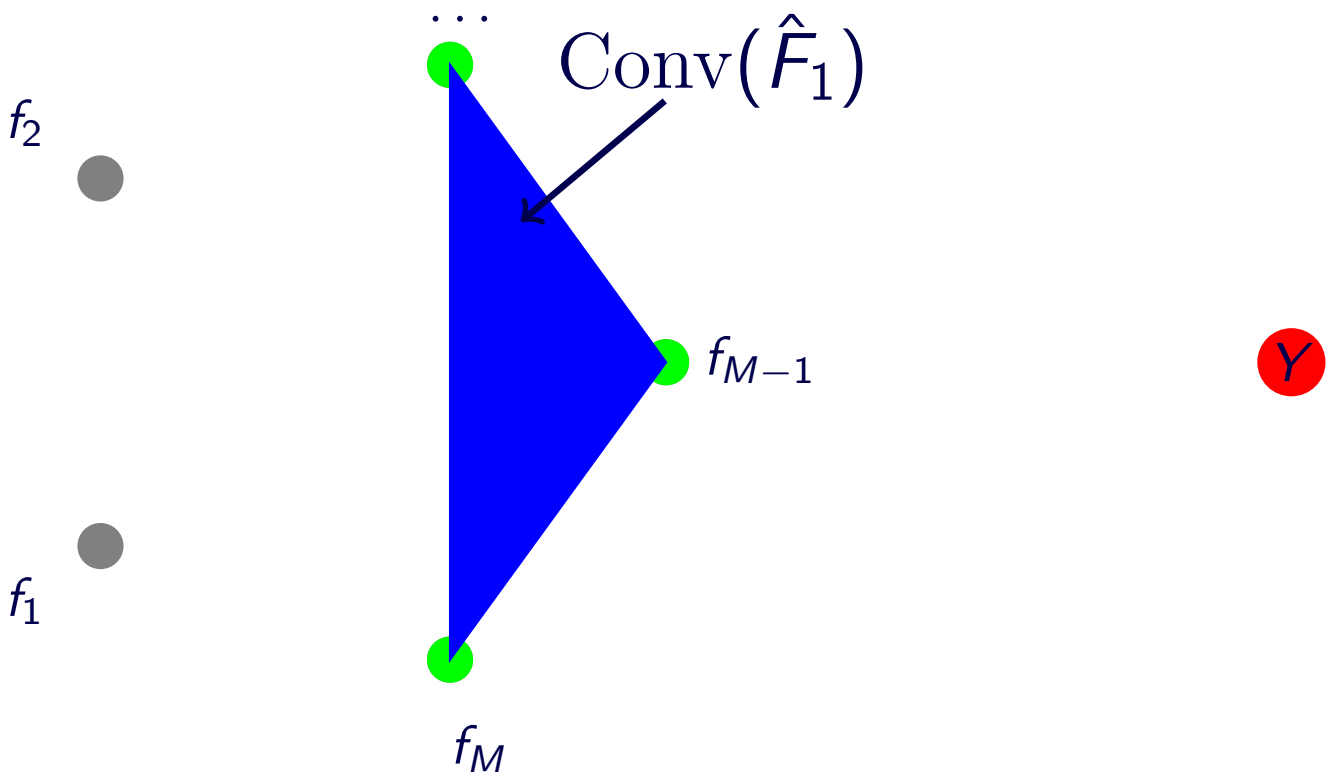

f_2
f_{M-1}
f_1
f_M


...
f_2
f_{M-1}
Y
f_1
f_M


...
f_2
f_{M-1}
Y
f_1
f_M


$\dots$
$\mathrm{Conv}(\hat{F}_1)$
f_2
f_{M-1}
Y
f_1
f_M

$$\tilde{f} \in \operatorname{Arg\,min}_{f \in \mathrm{Conv}(\hat{F}_1)} R_n(f)$$



Data splitting

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n}))$.

Data splitting

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \ldots, (X_{2n}, Y_{2n}))$.

$D_1 = ((X_1, Y_1), \ldots, (X_n, Y_n))$:
construction of the set $\hat{F}_1$ of approximatively good elements of F.

Data splitting

$2n$ observations :
$D_{2n} = ((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$.

$D_1 = ((X_1, Y_1), \dots, (X_n, Y_n))$:
construction of the set $\hat{F}_1$ of approximatively good elements of F.

$D_2 = ((X_{n+1}, Y_{n+1}), \dots, (X_{2n}, Y_{2n}))$:
construction of the ERM over $\mathrm{Conv}(\hat{F}_1)$

$$\tilde{f} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(\hat{F}_1)} \frac{1}{n} \sum_{i=n+1}^{2n} (f(X_i) - Y_i)^2.$$

Construction of $\hat{F}_1$ (preselection step)

1st step : ERM over F (over D_1) :

$$\hat{f} \in \operatorname{Arg}\min_{f \in F} R_n(f)$$

Construction of $\hat{F}_1$ (preselection step)

1st step : ERM over F (over D_1) :

$$\hat{f} \in \operatorname{Arg}\min_{f \in F} R_n(f)$$

2nd step : set of almost minimizers of the ER (over D_1) :

$$\hat{F}_1 = \left\{ f \in F : R_n(f) \leq R_n(\hat{f}) + C_1 \max(\alpha \|\hat{f} - f\|_{L_2^n}, \alpha^2) \right\},$$

where $\alpha = \sqrt{(x + \log M)/n}$, $x > 0$ is the confidence and

$$\|\hat{f} - f\|_{L_2^n}^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(X_i) - f(X_i))^2.$$

ERM over $\mathrm{Conv}(\hat{F}_1)$

$$\mathrm{Conv}(\hat{F}_1) = \Big\{ \sum_{f \in \hat{F}_1} \lambda_f f : \lambda_f \geq 0 \text{ and } \sum_{f \in \hat{F}_1} \lambda_f = 1 \Big\},$$

ERM over $\mathrm{Conv}(\hat{F}_1)$

$$\mathrm{Conv}(\hat{F}_1) = \Big\{ \sum_{f \in \hat{F}_1} \lambda_f f : \lambda_f \geq 0 \text{ and } \sum_{f \in \hat{F}_1} \lambda_f = 1 \Big\},$$

Construction of the ERM over $\mathrm{Conv}(F_1)$ with ER based on D_2 :

$$\tilde{f} \in \mathrm{Arg} \min_{f \in \mathrm{Conv}(\hat{F}_1)} \frac{1}{n} \sum_{i=n+1}^{2n} (f(X_i) - Y_i)^2.$$

Exact Oracle Inequality

Theorem

$\forall x > 0$, $\forall F$ of M functions, any target Y (all bounded by b), with ν^{2n}-probability at least $1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(1+x)\frac{\log M}{n}$$

Exact Oracle Inequality

Theorem

$\forall x > 0$, $\forall F$ of M functions, any target Y (all bounded by b), with ν^{2n}-probability at least $1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(1+x)\frac{\log M}{n}$$

Conclusion : $(\log M)/n$ is the optimal rate of aggregation and $\tilde{f}$ is an optimal aggregation procedure with confidence $2\exp(-x)$.

Exact Oracle Inequality

Theorem

$\forall x > 0$, $\forall F$ of M functions, any target Y (all bounded by b), with ν^{2n}-probability at least $1 - 2\exp(-x)$,

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(1+x)\frac{\log M}{n}$$

Conclusion : $(\log M)/n$ is the optimal rate of aggregation and $\tilde{f}$ is an optimal aggregation procedure with confidence $2\exp(-x)$.
Remark : This aggregation procedure is sparse in the sense that non-relevant elements in F have a zero-coefficient.

Model : agnostic functional learning

- Observations $(X_i, T(X_i))_{i=1,\dots,n}$ $(Y = T(X))$; T : Target function

Model : agnostic functional learning

- Observations $(X_i, T(X_i))_{i=1,\dots,n}$ $(Y = T(X))$; T : Target function
- $T \in \tau$; τ is convex ;

Model : agnostic functional learning

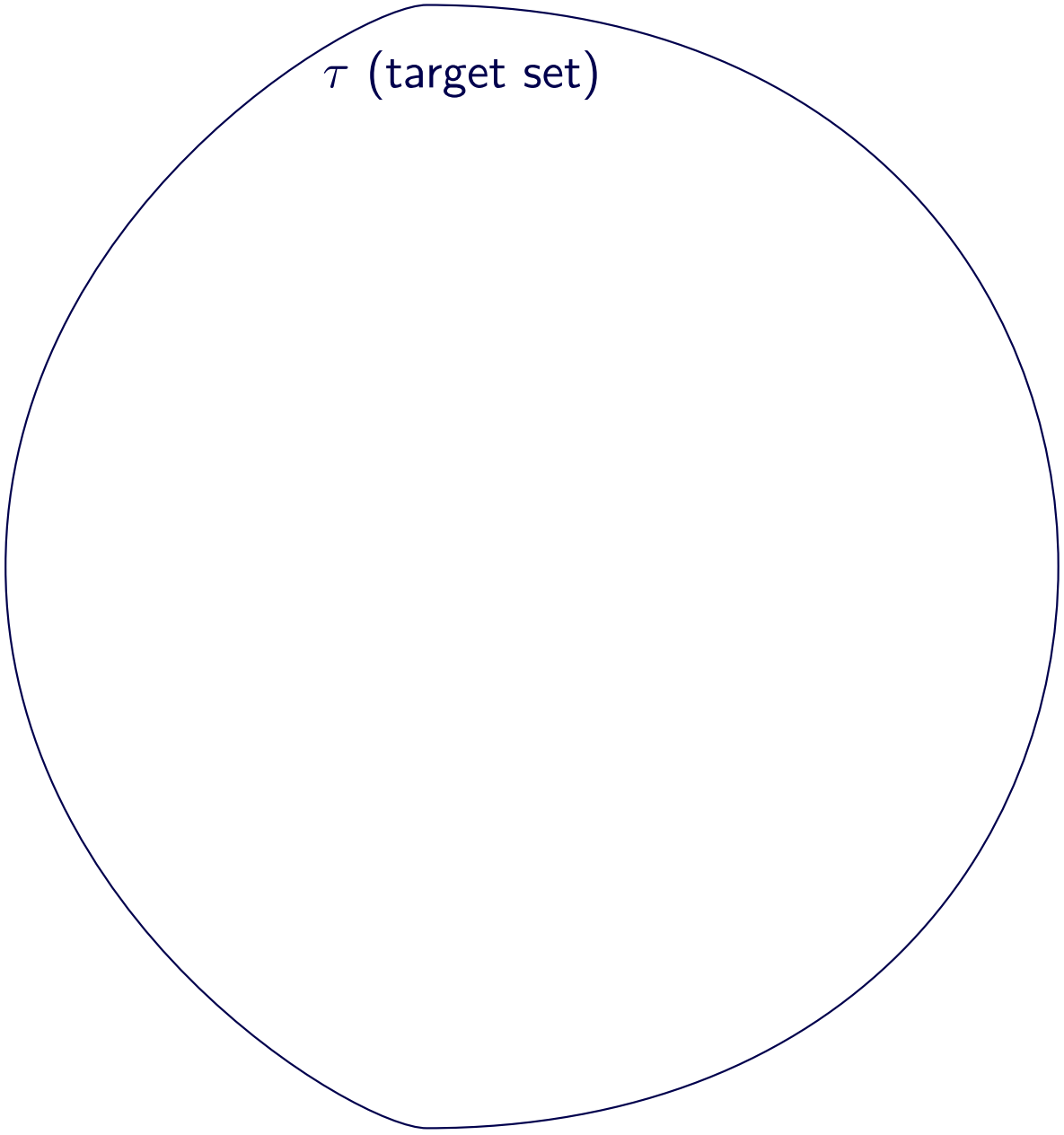
- Observations $(X_i, T(X_i))_{i=1,\dots,n}$ $(Y = T(X))$; T : Target function
- $T \in \tau$; τ is convex ;
- F (set of candidates) $\subset \tau$;

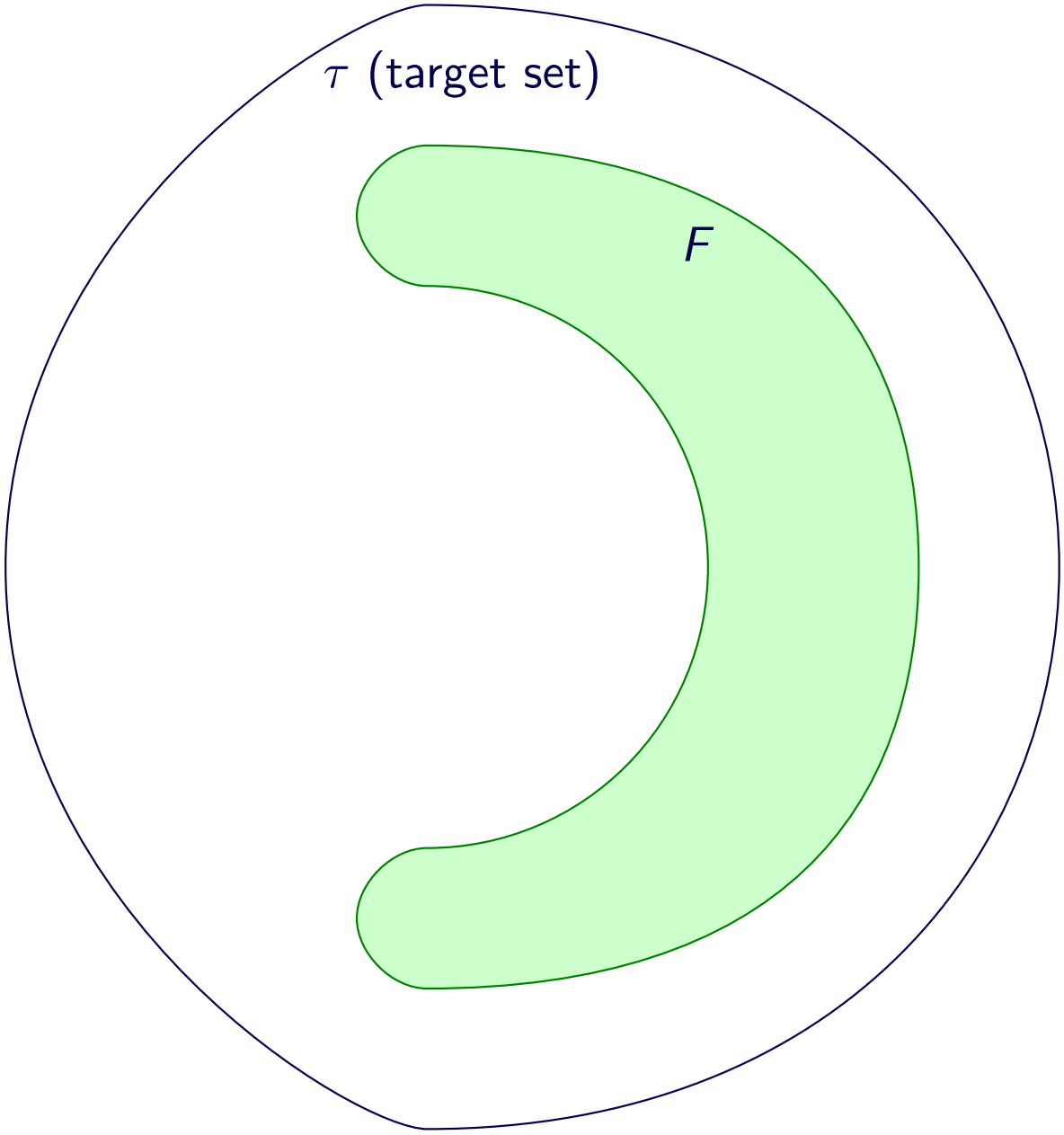
Model : agnostic functional learning

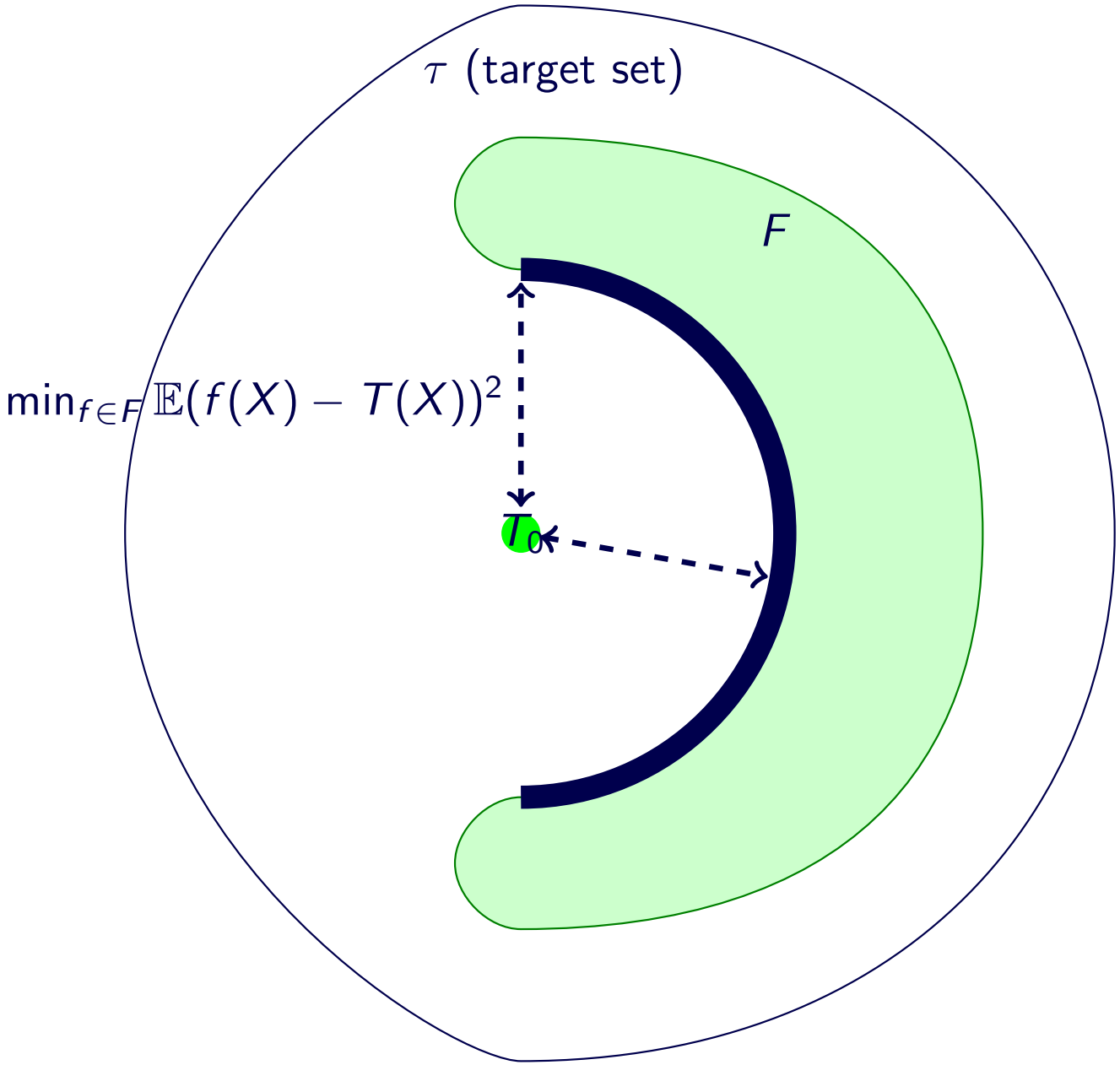
- Observations $(X_i, T(X_i))_{i=1,\dots,n}$ $(Y = T(X))$; T : Target function
- $T \in \tau$; τ is convex ;
- F (set of candidates) $\subset \tau$;

agnostic functional learning problem : Find a procedure which has a risk as close as possible to

$$\inf_{f \in F} \mathbb{E}(f(X) - T(X))^2$$


τ (target set)


τ (target set)
F


τ (target set)
F
$\min_{f \in F} \mathbb{E}(f(X) - T(X))^2$
T_0

Empirical risk minimization algorithm

Exact Oracle Inequality : Construction of a procedure $\hat{f}(\cdot, \mathcal{D})$ such that

$$\mathbb{E}(\hat{f}(X) - T(X))^2 \leq \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2 + r_n(F)$$

Empirical risk minimization algorithm

Exact Oracle Inequality : Construction of a procedure $\hat{f}(\cdot, \mathcal{D})$ such that

$$\mathbb{E}(\hat{f}(X) - T(X))^2 \leq \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2 + r_n(F)$$

Empirical risk minimization algorithm :

$$\hat{f} \in \operatorname{Arg}\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - T(X_i))^2$$

Upper bound

Symmetrization

Upper bound

Symmetrization+majoration of Rademacher processes by Gaussian processes

Upper bound

Symmetrization+majoration of Rademacher processes by Gaussian processes+Dudley's entropy integral :

Upper bound

Symmetrization+majoration of Rademacher processes by Gaussian processes+Dudley's entropy integral :

$$\mathbb{E}(\hat{f}(X) - T(X))^2 - \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2$$
$$\lesssim \frac{1}{\sqrt{n}} \mathbb{E}_X \int_0^{\mathrm{diam}(P_\sigma \mathcal{F}, |\cdot|_{2,n})} \sqrt{\mathcal{N}(P_\sigma \mathcal{F}, |\cdot|_{2,n}, \epsilon)} d\epsilon$$

Upper bound

Symmetrization+majoration of Rademacher processes by Gaussian processes+Dudley's entropy integral :

$$\mathbb{E}(\hat{f}(X) - T(X))^2 - \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2$$
$$\lesssim \frac{1}{\sqrt{n}} \mathbb{E}_X \int_0^{\mathrm{diam}(P_\sigma \mathcal{F}, |\cdot|_{2,n})} \sqrt{\mathcal{N}(P_\sigma \mathcal{F}, |\cdot|_{2,n}, \epsilon)} d\epsilon$$

$$\mathcal{F} = \{(f - T)^2 - (f^* - T)^2 : f \in F\} \quad \text{(excess loss class)}$$
$$P_\sigma \mathcal{F} = \{(g(X_1), \ldots, g(X_n)) : g \in \mathcal{F}\} \quad \text{(coordinate projection)}$$
$$|u|_{2,n} = \sqrt{\frac{1}{n} \sum_{i=1}^n u_i^2} \quad \text{(normalized } l_2^n\text{-norm)}$$

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm

$\exists T \in \tau$ s.t.

$$\mathbb{E}(\hat{f}(X) - T(X))^2 - \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2 \geq r_n(F')$$

where $F' \subset F$.

Aim

Lower bound for the excess risk of the Empirical risk minimization algorithm

$\exists T \in \tau$ s.t.

$$\mathbb{E}(\hat{f}(X) - T(X))^2 - \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2 \geq r_n(F')$$

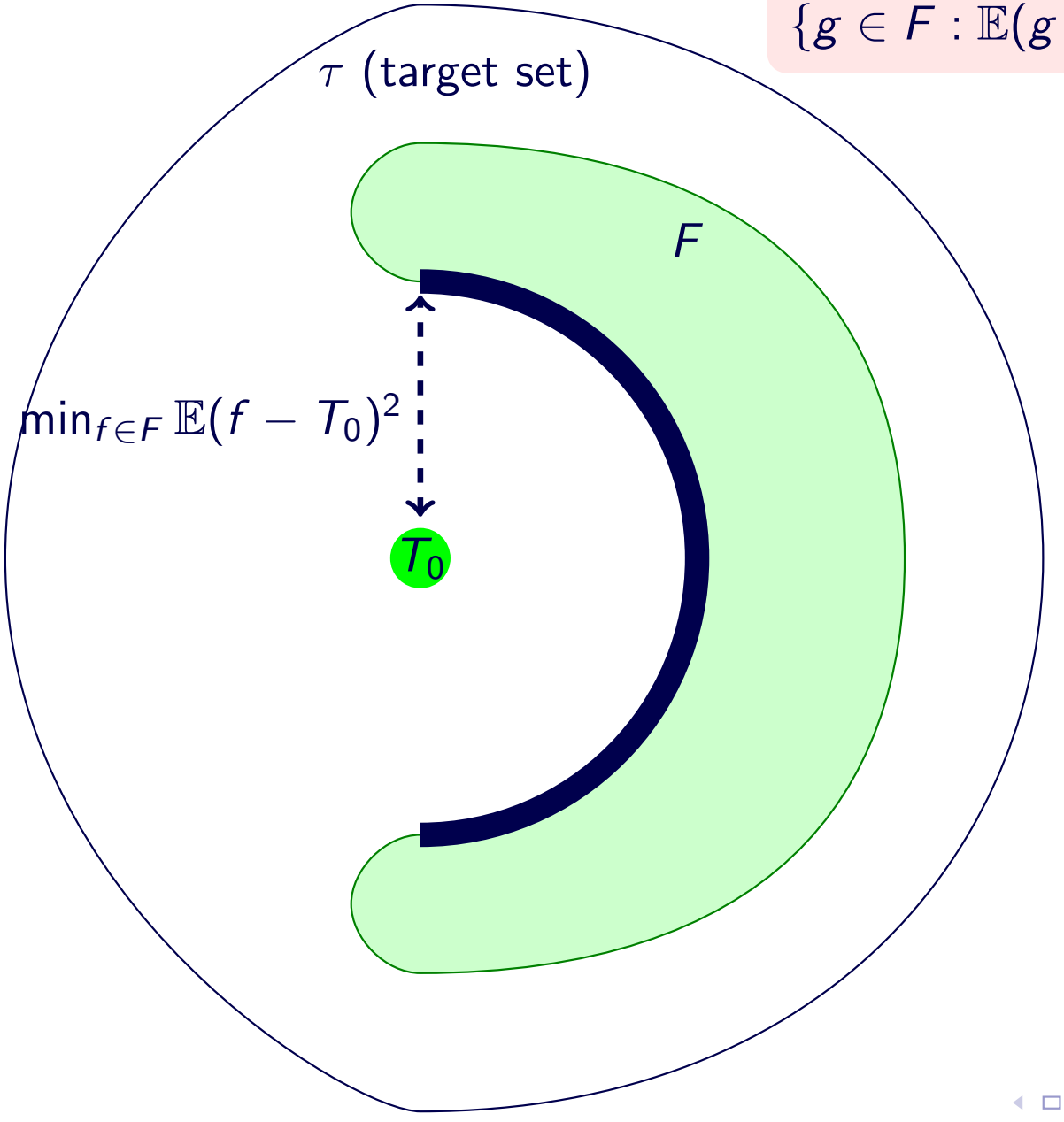
where $F' \subset F$.

Assumption :

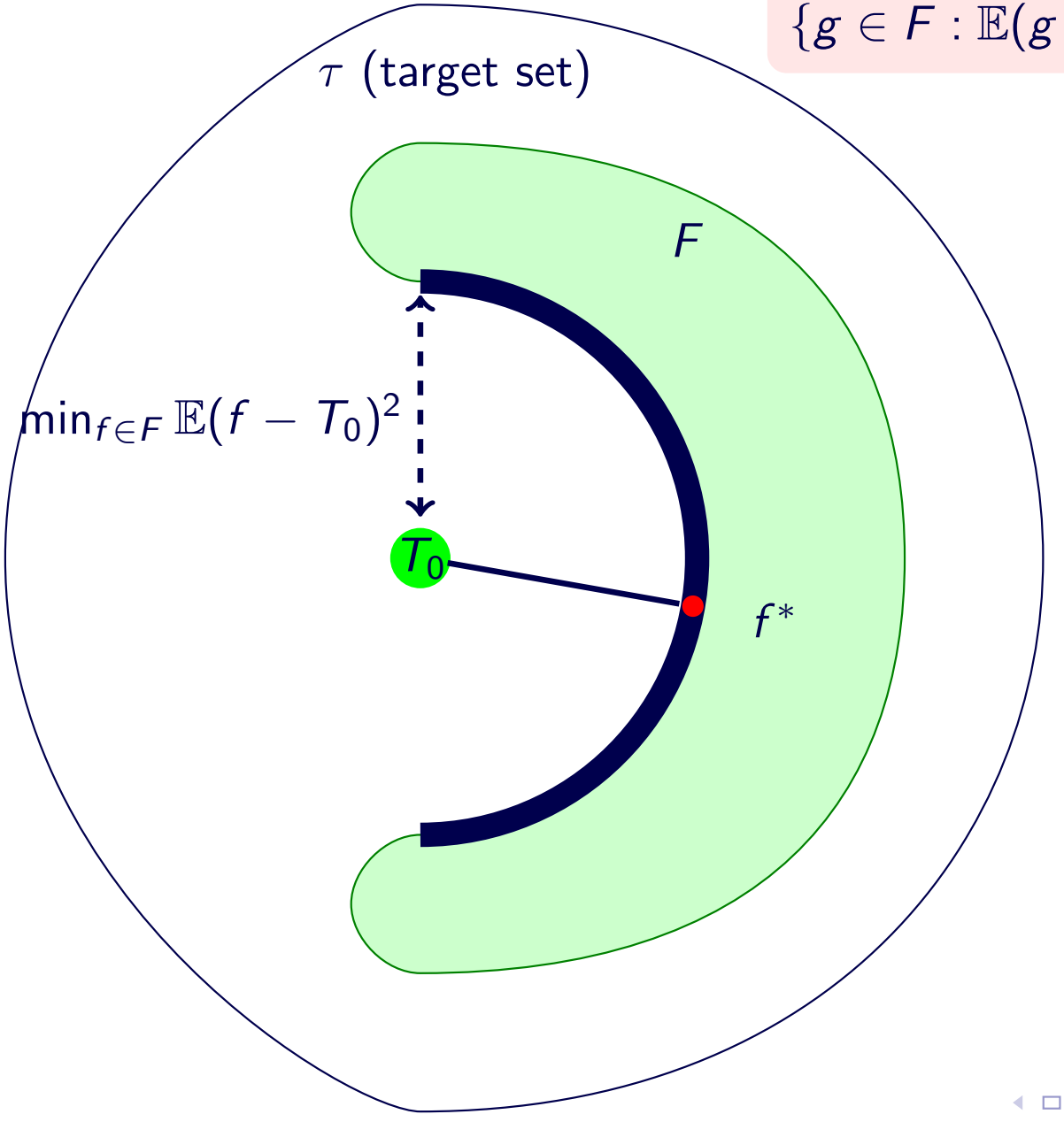
$\exists T_0 \in \tau$ s.t. $\mathrm{card}(\mathcal{M}(T_0)) \geq 2$ where

$$\mathcal{M}(T) = \{f \in F : \mathbb{E}(f(X) - T(X))^2 = \inf_{f \in F} \mathbb{E}(f(X) - T(X))^2\}$$

$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$

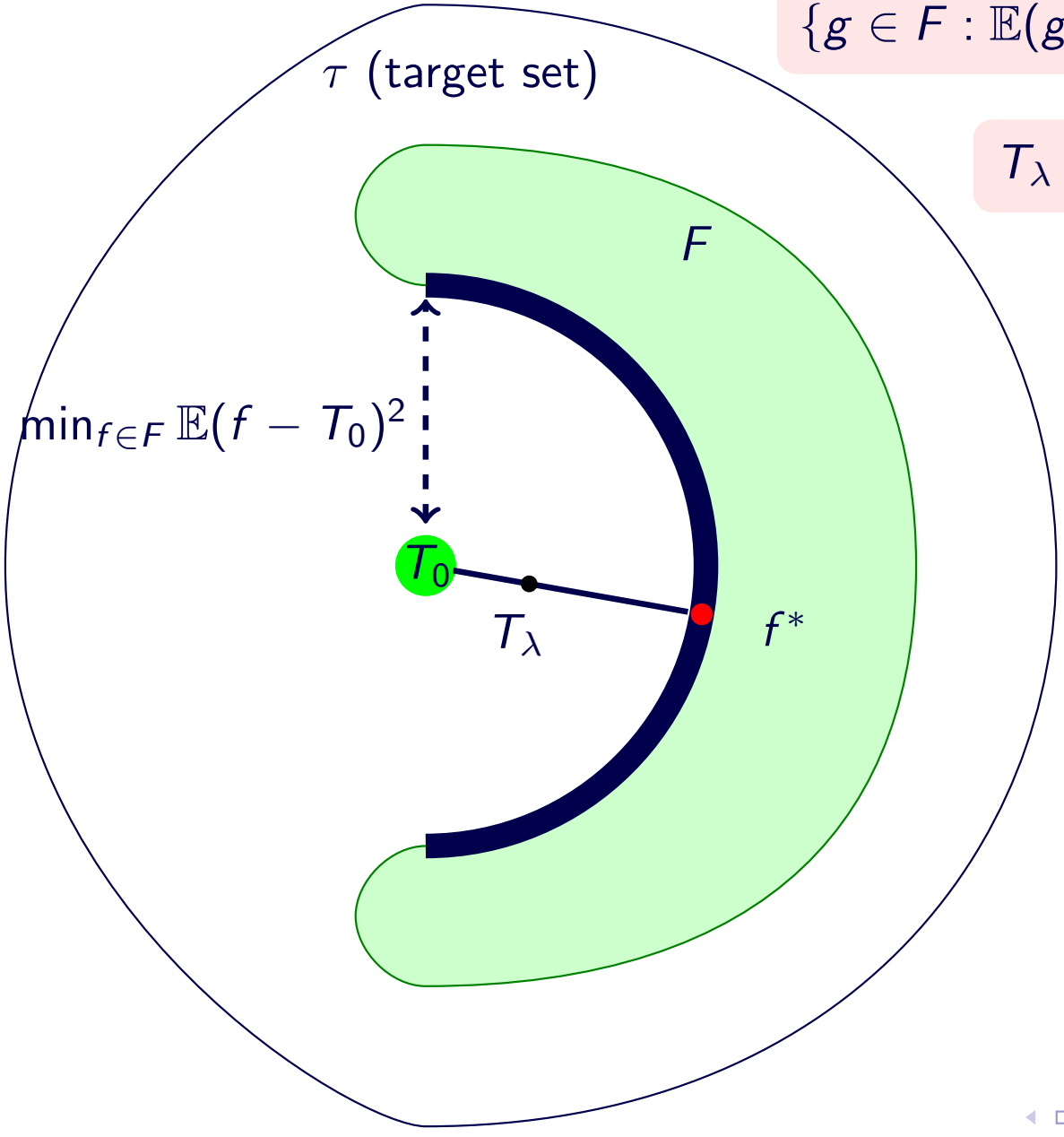


$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$



$$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$$

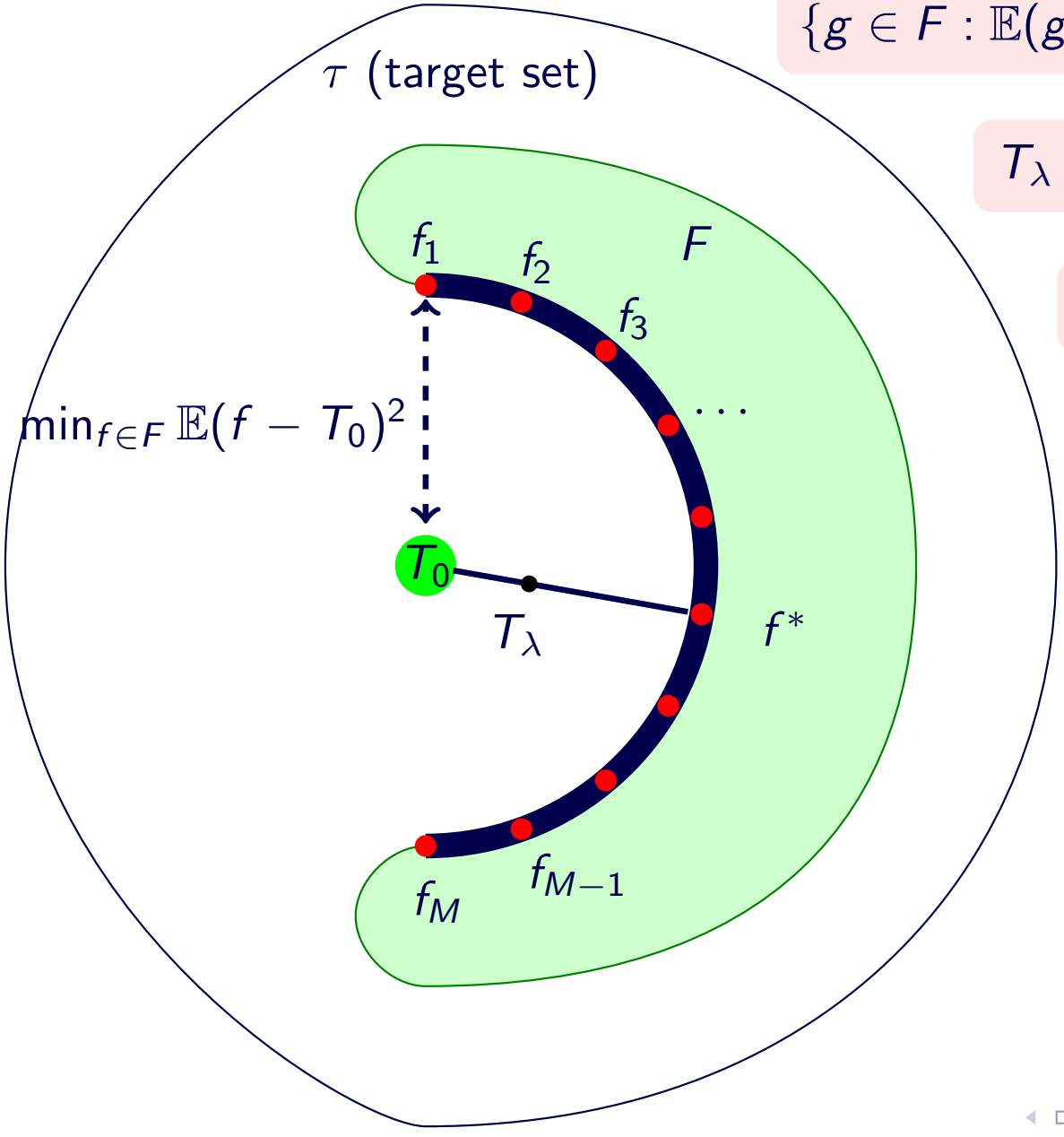
$$T_\lambda = (1 - \lambda) T_0 + \lambda f^*$$



$$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$$

$$T_\lambda = (1 - \lambda) T_0 + \lambda f^*$$

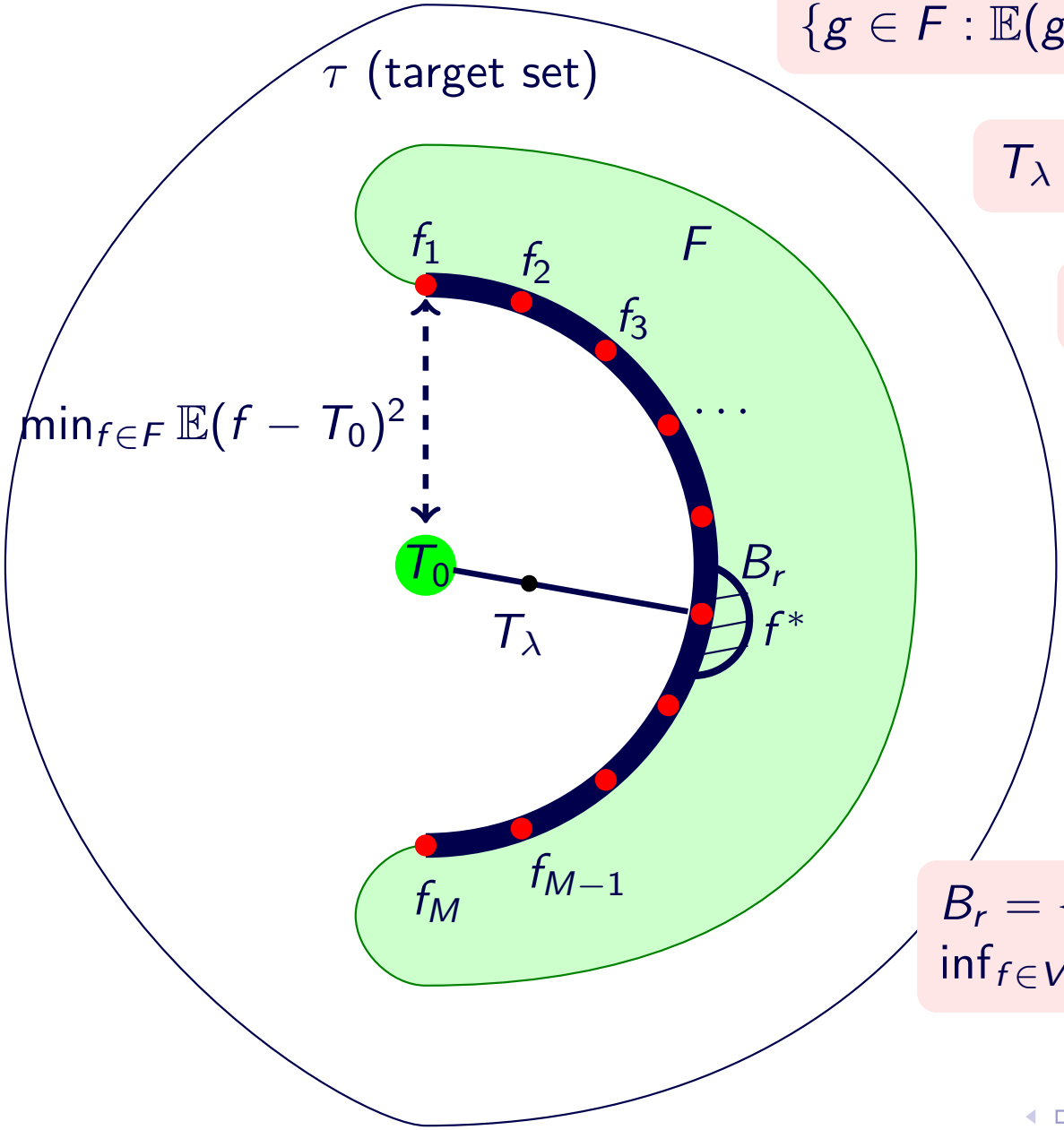
$$V_m = \{f_1, \dots, f_M\}$$



$$\{g \in F : \mathbb{E}(g - T_0)^2 = \inf_F \mathbb{E}(f - T_0)^2\}$$

$$T_\lambda = (1-\lambda)T_0 + \lambda f^*$$

$$V_m = \{f_1, \dots, f_M\}$$



$$B_r = \{f : R(f) \le R(f^*) + r\}$$
$$\inf_{f\in V_m} R_n^\lambda(f) < \inf_{f\in B_r} R_n^\lambda(f)$$

Outline of the proof

The core of the proof is to find a set Q that can "compete" with $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ in the sense that the empirical excess risk function

$$\mathcal{E}_n : f \in F \longmapsto \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\lambda(f)(X_i) = P_n\mathcal{L}_\lambda(f)$$

will be more negative on Q than on it can possibly be on B_r $(\mathcal{L}_\lambda(f) := (f - T_\lambda)^2 - (f^* - T_\lambda)^2)$.

Outline of the proof

The core of the proof is to find a set Q that can "compete" with $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ in the sense that the empirical excess risk function

$$\mathcal{E}_n : f \in F \longmapsto \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\lambda(f)(X_i) = P_n\mathcal{L}_\lambda(f)$$

will be more negative on Q than on it can possibly be on B_r ($\mathcal{L}_\lambda(f) := (f - T_\lambda)^2 - (f^* - T_\lambda)^2$).
Thus, the ERM $\hat{f}_\lambda \notin B_r$, and thus, with a certain probability,

$$\mathbb{E}\big[\mathcal{L}_\lambda(\hat{f}_\lambda)|D\big] > r.$$

Proof in two parts :

- $\mathcal{E}_n$ is likely to be very negative on $\{f \in F : \mathbb{E}\mathcal{L}(f) = \min_{f\in F} \mathbb{E}\mathcal{L}(f)\}$;
- find some r on which the oscillations of $\mathcal{E}_n$ in B_r are small.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q \in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1 \leq i,j \leq N})$.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \dots, q_N \in Q$, $(G_{q_1}, \dots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \dots, q_N \in Q$, $(G_{q_1}, \dots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \in A(t_1, \ldots, t_d)\Big) - \mathbb{P}(G \in A(t_1, \ldots, t_d))\right| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \ldots, q_N \in Q$, $(G_{q_1}, \ldots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\leq i,j\leq N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \in A(t_1, \ldots, t_d)\Big) - \mathbb{P}(G \in A(t_1, \ldots, t_d))\right| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

where $A(t_1, \ldots, t_d) = \{v = (v_1, \ldots, v_d) \in \mathbb{R}^d : x_j \leq t_j, \forall j\}$

Gaussian process end multidimensional CLT

Gaussian process : Let $Q \subset L_2(\mu)$. $(G_q)_{q\in Q}$ is a *Canonical Gaussian process* associated with Q when $\forall N \in \mathbb{N}, \forall q_1, \dots, q_N \in Q$, $(G_{q_1}, \dots, G_{q_N}) \sim \mathcal{N}_N(0, (\langle q_i, q_j \rangle)_{1\le i,j\le N})$.
A common measure of the "complexity" of Q is

$$H(Q) := \mathbb{E} \sup_{q\in Q} G_q$$

Multidimensional CLT : $(V_i)_{i\in\mathbb{N}}$: sequence of d-dimensional i.i.d.r.v. with zero mean and finite L_2-norm.

$$\left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i \in A(t_1, \dots, t_d)\Big) - \mathbb{P}(G \in A(t_1, \dots, t_d))\right| \longrightarrow 0 \text{ as } n \longrightarrow +\infty,$$

where $A(t_1, \dots, t_d) = \{v = (v_1, \dots, v_d) \in \mathbb{R}^d : x_j \le t_j, \forall j\}$ and G is a d-dimensional Gaussian process with zero mean and covariance matrix $(\mathbb{E}V^{(i)}V^{(j)})_{1\le i,j\le d}$.

Multivariate CLT outside B_r

Fix a finite set $Q' \subset Q := \{\mathcal{L}(f) : \mathbb{E}\mathcal{L}(f) = \min_{f \in F} \mathbb{E}\mathcal{L}(f)\}$ for which $H(Q') \geq H(Q)/2$ and $0 \in Q'$.

Multivariate CLT outside B_r

Fix a finite set $Q' \subset Q := \{\mathcal{L}(f) : \mathbb{E}\mathcal{L}(f) = \min_{f \in F} \mathbb{E}\mathcal{L}(f)\}$ for which $H(Q') \geq H(Q)/2$ and $0 \in Q'$.

Theorem

$\exists n_0 = n_0(Q')$ s.t. $\forall n \geq n_0$, with μ^n-probability at least c_1,

$$\inf_{\mathcal{L}(f) \in Q'} P_n \mathcal{L}_{\lambda_n}(f) \leq -c_2 \frac{H(Q)}{\sqrt{n}}$$

where $\lambda_n = c_3 H(Q)/\sqrt{n}$.

Uniform Central Limit Theorem

Recall that a bounded class of functions F is μ-Donsker if and only if for $\forall u > 0, \exists \delta > 0, \exists n_0$ s.t. $\forall n \geq n_0, \mathrm{osc}_n(F, \delta) \leq u$ where

$$\mathrm{osc}_n(F, \delta) = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{\{f,h \in F : \|f-h\| \leq \delta\}} \left| \sum_{i=1}^{n} g_i (f-h)(X_i) \right|,$$

where $g_1, \ldots, g_n$ are n i.i.d. standard Gaussian variables.

Uniform Central Limit Theorem

Recall that a bounded class of functions F is μ-Donsker if and only if for $\forall u > 0, \exists \delta > 0, \exists n_0$ s.t. $\forall n \geq n_0, \mathrm{osc}_n(F, \delta) \leq u$ where

$$\mathrm{osc}_n(F, \delta) = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{\{f,h \in F : \|f-h\| \leq \delta\}} \left| \sum_{i=1}^{n} g_i (f-h)(X_i) \right|,$$

where $g_1, \ldots, g_n$ are n i.i.d. standard Gaussian variables.

δ s.t. $\forall n \geq N(F)$, $\mathrm{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$.

UCLT around f^*

Now we are ready to control the oscillation of the empirical excess risk function uniformly over the set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda \leq r\}$.

UCLT around f^*

Now we are ready to control the oscillation of the empirical excess risk function uniformly over the set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda \leq r\}$.

Theorem

$\exists c_3$ s.t. $\forall n \geq n_1$, with μ^n-probability at least $1 - c_1/2$,

$$\inf_{\{f \in F : \mathbb{E}\mathcal{L}_{\lambda_n}(f) \leq r_n\}} P_n \mathcal{L}_{\lambda_n}(f) \geq -\frac{c_2 H(Q)}{2\sqrt{n}}$$

where

$$r_n = c_3 \frac{H(Q)}{\sqrt{n}} \delta^2 \| T - f^* \|^2$$

Main Theorem

Theorem

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$.

Main Theorem

Theorem

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$. Set $Q = \{\mathcal{L}(f) : f \in F,\ \mathbb{E}\mathcal{L}(f) = 0\}$.

Main Theorem

Theorem

Let $F \subset B(L_\infty)$ which is μ-Donsker and assume that $\exists T_0 \in \tau$ s.t. $\mathcal{M}(T_0) \geq 2$. Set $Q = \{\mathcal{L}(f) : f \in F,\ \mathbb{E}\mathcal{L}(f) = 0\}$.
$\exists C_1, C_2, N(F)$ s.t. $\forall n \geq N(F)$, with μ^n-probability at least C_1,

$$\mathbb{E}\mathcal{L}_{\lambda_n}(\hat{f}_{\lambda_n}) \geq C_2 \frac{H(Q)}{\sqrt{n}} \delta^2 \| T - f^* \|$$

where δ is s.t. $\forall n \geq N(F)$, $\mathrm{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$ and $\lambda_n = C_2 H(Q)/\sqrt{n}$.

Subgaussian regression

$$Y = f_0(X) + \sigma\varepsilon$$

$f_0 : \mathbb{R}^d \to \mathbb{R}$; ε : noise (noise level σ is known)

Subgaussian regression

$$Y = f_0(X) + \sigma\varepsilon$$

$f_0 : \mathbb{R}^d \to \mathbb{R}$; ε : noise (noise level σ is known)

n observations :

$$D_n := [(X_i, Y_i) \; ; \; 1 \leq i \leq n],$$

Notation : $P_n = P[\cdot | X_1, \ldots, X_n]$ and E_n : expectation w.r.t. P_n.

Assumptions

1. ε is centered and subgaussian
   $(\exists b > 0; E_n[\exp(t\varepsilon)] \leq \exp(b^2 t^2/2), \forall t > 0)$.

Assumptions

1. ε is centered and subgaussian
   $(\exists b > 0; E_n[\exp(t\varepsilon)] \leq \exp(b^2 t^2/2), \forall t > 0)$.

2. X has a compact support, (we do not need $P_X << \lambda_d$). Take $\mathrm{Supp}(P_X) \subset [0,1]^d$.

Assumptions

1. ε is centered and subgaussian
   $(\exists b > 0; E_n[\exp(t\varepsilon)] \leq \exp(b^2 t^2/2), \forall t > 0)$.

2. X has a compact support, (we do not need $P_X << \lambda_d$). Take $\mathrm{Supp}(P_X) \subset [0,1]^d$.

3. $f_0 \in \mathcal{F}$

Complexity assumption

1. $\mathcal{F}$ is endowed with a semi-norm $|\cdot|_{\mathcal{F}}$.

Complexity assumption

1. $\mathcal{F}$ is endowed with a semi-norm $|\cdot|_{\mathcal{F}}$.

2. $\mathcal{F} \subset C([0,1]^d)$

Complexity assumption

1. $\mathcal{F}$ is endowed with a semi-norm $|\cdot|_{\mathcal{F}}$.

2. $\mathcal{F} \subset C([0,1]^d)$

3. (C_β) $\exists \beta \in (0,2), D > 0$ s.t. $\forall \delta, R > 0$

$$H_\infty\big(\delta, \mathcal{F}(R)\big) \leq D(R/\delta)^\beta$$

$$\mathcal{F}(R) := \{f \in \mathcal{F} : |f|_{\mathcal{F}} \leq R\}$$

$$H_\infty(\delta, \mathcal{F}(R)) = \log \min \big(N \in \mathbb{N} : \mathcal{F}(R) \subset \cup_{j=1}^N \mathcal{B}_\infty(f_j, \delta)\big).$$

example : Anisotropic Besov space

1. $\mathbf{s} = (s_1, \ldots, s_d)$ with $s_i > 0$: vector of directional smoothness (s_i smoothness in direction e_i)

example : Anisotropic Besov space

1. $\mathbf{s} = (s_1, \ldots, s_d)$ with $s_i > 0$: vector of directional smoothness (s_i smoothness in direction e_i)
2. $f \in L^p(\mathbb{R}^d)$ belongs to the anisotropic Besov space $B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ if the semi-norm

$$|f|_{B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)} := \sum_{i=1}^{d} \left(\int_0^1 (t^{-s_i} \left\| \Delta^{k_i}_{te_i} f \right\|_p)^q \frac{dt}{t} \right)^{1/q} < +\infty$$

$\Delta^1_h f(x) = f(x+h) - f(x)$ and $\Delta^k_h f = \Delta^1_h(\Delta^{k-1}_h f)(x)$ and $k_i > s_i$.

example : Anisotropic Besov space

1. $\mathbf{s} = (s_1, \ldots, s_d)$ with $s_i > 0$: vector of directional smoothness (s_i smoothness in direction e_i)
2. $f \in L^p(\mathbb{R}^d)$ belongs to the anisotropic Besov space $B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ if the semi-norm

$$|f|_{B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)} := \sum_{i=1}^{d} \left(\int_0^1 (t^{-s_i} \left\| \Delta^{k_i}_{te_i} f \right\|_p)^q \frac{dt}{t} \right)^{1/q} < +\infty$$

   $\Delta^1_h f(x) = f(x+h) - f(x)$ and $\Delta^k_h f = \Delta^1_h(\Delta^{k-1}_h f)(x)$ and $k_i > s_i$.
3. $\mathbf{s} = (s, \ldots, s) \Longrightarrow B^{\mathbf{s}}_{p,q}$ is the standard isotropic Besov space.

example : Anisotropic Besov space

$\Omega \subset \mathbb{R}^d$: arbitrary domain. $B^{\mathbf{s}}_{p,q}(\Omega)$: set of all $f \in L^p(\Omega)$ such that $\exists g \in B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ with restriction $g|\Omega$ to Ω equal to f in $L^p(\Omega)$. Moreover,

$$\|f\|_{B^{\mathbf{s}}_{p,q}(\Omega)} = \inf_{g : g|\Omega = f} \|g\|_{B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)},$$

where the infimum is taken over all $g \in B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ such that $g|\Omega = f$.

example : Anisotropic Besov space

$\Omega \subset \mathbb{R}^d$: arbitrary domain. $B^{\mathbf{s}}_{p,q}(\Omega)$: set of all $f \in L^p(\Omega)$ such that $\exists g \in B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ with restriction $g|\Omega$ to Ω equal to f in $L^p(\Omega)$. Moreover,

$$\|f\|_{B^{\mathbf{s}}_{p,q}(\Omega)} = \inf_{g : g|\Omega = f} \|g\|_{B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)},$$

where the infimum is taken over all $g \in B^{\mathbf{s}}_{p,q}(\mathbb{R}^d)$ such that $g|\Omega = f$.

Theorem (Birman and Solomjak (67))

Let $1 \leq p, q \leq \infty$ and $\mathbf{s} = (s_1, \ldots, s_d)$ where $s_i > 0$, and let s be the harmonic mean of $\mathbf{s}$

$$\frac{1}{s} = \frac{1}{d} \sum_{j=1}^{d} \frac{1}{s_j}$$

If $s > d/p$, then

$$B^{\mathbf{s}}_{p,q}(\Omega) \subset C(\Omega),$$

$$H_\infty(\delta, B^{\mathbf{s}}_{p,q}(R)) \leq D(R/\delta)^{s/d}, \forall \delta, R > 0$$

Estimation : PERM

Definition (PERM)

Let $\lambda = (h, \mathcal{F})$ be fixed. We say that $\bar{f}_\lambda$ is a penalized empirical risk minimizer if it minimizes

$$R_n(f) + \mathrm{pen}(f)$$

over $\mathcal{F}$, where $\mathrm{pen}(f) := h^2 |f|_{\mathcal{F}}^{\alpha}$ for some $\alpha > 0$ and where

$$R_n(f) := \|Y - f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

is the empirical risk of f over the sample D_n.

Non-adaptive rate of convergence

Theorem

$\mathcal{F}$ satisfying (C_β),

$$h = an^{-1/(2+\beta)}$$

and $\alpha > 2\beta/(\beta+2)$.

Non-adaptive rate of convergence

Theorem

$\mathcal{F}$ satisfying (C_β),

$$h = an^{-1/(2+\beta)}$$

and $\alpha > 2\beta/(\beta+2)$.

1. $$E_n\|\bar{f}_\lambda - f_0\|_n^2 \leq C_1(1 + |f_0|_{\mathcal{F}}^\alpha)n^{-2/(2+\beta)}$$

Non-adaptive rate of convergence

Theorem

$\mathcal{F}$ satisfying (C_β),

$$h = an^{-1/(2+\beta)}$$

and $\alpha > 2\beta/(\beta+2)$.

1. $$E_n\|\bar{f}_\lambda - f_0\|_n^2 \leq C_1(1 + |f_0|_{\mathcal{F}}^\alpha)n^{-2/(2+\beta)}$$

2. *If $\|\bar{f}_\lambda - f_0\|_\infty \leq Q$ and $\|f_0\|_\infty \leq Q$ then*

$$E^n\|\bar{f}_\lambda - f_0\|^2 \leq C_2(1 + |f_0|_{\mathcal{F}}^\alpha)n^{-2/(2+\beta)}$$

Rate of convergence for anisotropic Besov space

Corollary

$\mathcal{F} = B^{\mathbf{s}}_{p,\infty}$ and $h = an^{-s/(2s+d)}$ where s is the harmonic mean of $\mathbf{s}$ such that $s > d/p$ and $\|\bar{f}_\lambda - f_0\|_\infty, \|f_0\|_\infty \leq Q$. Then, we have :

Rate of convergence for anisotropic Besov space

Corollary

$\mathcal{F} = B^{\mathbf{s}}_{p,\infty}$ and $h = an^{-s/(2s+d)}$ where s is the harmonic mean of $\mathbf{s}$ such that $s > d/p$ and $\|\bar{f}_\lambda - f_0\|_\infty, \|f_0\|_\infty \leq Q$. Then, we have :

$$E\|\bar{f}_\lambda - f_0\|^2 \leq C_3(1 + |f_0|^2_{B^{\mathbf{s}}_{p,\infty}})n^{-2s/(2s+d)}.$$

Problem of adaptation

Case $\mathcal{F} = B^{\mathbf{s}}_{p,\infty}(\Omega)$.

Problem of adaptation

Case $\mathcal{F} = B^{\mathbf{s}}_{p,\infty}(\Omega)$.

$$h = an^{-s/(2s+d)}$$

depends on s the harmonic mean of $\mathbf{s}$

Problem of adaptation

Case $\mathcal{F} = B^{\mathbf{s}}_{p,\infty}(\Omega)$.

$$h = an^{-s/(2s+d)}$$

depends on s the harmonic mean of $\mathbf{s}$ which is unknown

$$\Downarrow$$

Problem of adaptation

Problem of adaptation

Case $\mathcal{F} = B^{\mathbf{s}}_{p,\infty}(\Omega)$.

$$h = an^{-s/(2s+d)}$$

depends on s the harmonic mean of $\mathbf{s}$ which is unknown

$$\Downarrow$$

Problem of adaptation

We want to construct a procedure $\tilde{f}$

- independent of s

Problem of adaptation

Case $\mathcal{F} = B^{\mathbf{s}}_{p,\infty}(\Omega)$.

$$h = an^{-s/(2s+d)}$$

depends on s the harmonic mean of $\mathbf{s}$ which is unknown

$$\Downarrow$$

Problem of adaptation

We want to construct a procedure $\tilde{f}$

- independent of s
- if $f \in B^{\mathbf{s}}_{p,\infty}$ with s the harmonic mean of $\mathbf{s}$:

$$E\|\tilde{f} - f_0\|^2 \leq C_3(1 + |f_0|^2_{B^{\mathbf{s}}_{p,\infty}})n^{-2s/(2s+d)}.$$

Aggregation of PERM

1. Split of the data

Aggregation of PERM

1. Split of the data
2. $D_1 = ((X_1, Y_1), \ldots, (X_m, Y_m))$: construction of $\bar{f}_s$ the PERM with $h = an^{-s/(2s+d)}$ for different $s \in \mathcal{S}$:

$$\mathcal{S} = \left\{ b + \frac{k}{\log n} : k = 0, \ldots, B \log n \right\}.$$

Aggregation of PERM

1. Split of the data
2. $D_1 = ((X_1, Y_1), \dots, (X_m, Y_m))$: construction of $\bar{f}_s$ the PERM with $h = an^{-s/(2s+d)}$ for different $s \in \mathcal{S}$:

$$\mathcal{S} = \Big\{ b + \frac{k}{\log n} : k = 0, \dots, B \log n \Big\}.$$

3. $D_2 = ((X_{m+1}, Y_{m+1}), \dots, (X_n, Y_n))$: construction of the aggregation method

$$\tilde{f} \in \operatorname{Arg} \min_{f \in \operatorname{Conv}(\hat{F}_1)} R^{(2)}_{(n-m)/2}(f),$$

where the dictionary is $F = \{\bar{f}_s \mathbb{1}_{|\bar{f}_s| \le Q}, s \in \mathcal{S}\}$.

Result

Theorem

$\forall \mathbf{s} \in \mathbb{R}^d$ *s.t. the harmonic mean* s *is s.t.* $s > d/p$ *and* $s \in [b, B]$, *then* $\forall f_0 \in B^{\mathbf{s}}_{p,\infty} \cap \mathcal{B}_\infty(Q)$,

Result

Theorem

$\forall \mathbf{s} \in \mathbb{R}^d$ *s.t. the harmonic mean* s *is s.t.* $s > d/p$ *and* $s \in [b, B]$, *then* $\forall f_0 \in B^{\mathbf{s}}_{p,\infty} \cap \mathcal{B}_\infty(Q)$,

$$E^n \|\tilde{f} - f_0\|^2 \leq C_3 (1 + |f_0|^2_{B^{\mathbf{s}}_{p,\infty}}) n^{-2s/(2s+d)}.$$

Result

Theorem

$\forall \mathbf{s} \in \mathbb{R}^d$ *s.t. the harmonic mean* s *is s.t.* $s > d/p$ *and* $s \in [b, B]$, *then* $\forall f_0 \in B^{\mathbf{s}}_{p,\infty} \cap \mathcal{B}_\infty(Q)$,

$$E^n \|\tilde{f} - f_0\|^2 \leq C_3 (1 + |f_0|^2_{B^{\mathbf{s}}_{p,\infty}}) n^{-2s/(2s+d)}.$$

$\tilde{f}$ adapts automatically to the "regularity parameter" s