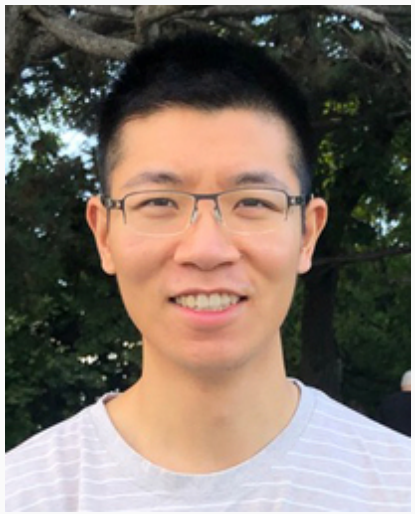
Evaluating algorithms and model classes: fundamental limits in the distribution-free setting

Rina Foygel Barber (University of Chicago)

http://rinafb.github.io/

Collaborators



Yuetian Luo
(Rutgers)

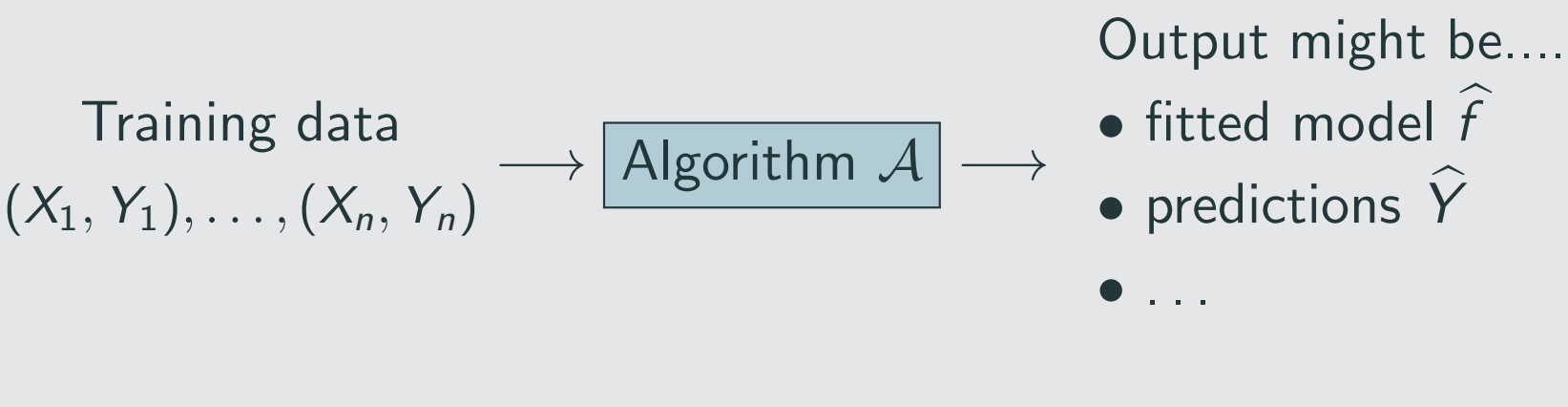
Manuel Müller
(Cambridge)

Luo & B. 2024, *The Limits of Assumption-free Tests for Algorithm Performance*

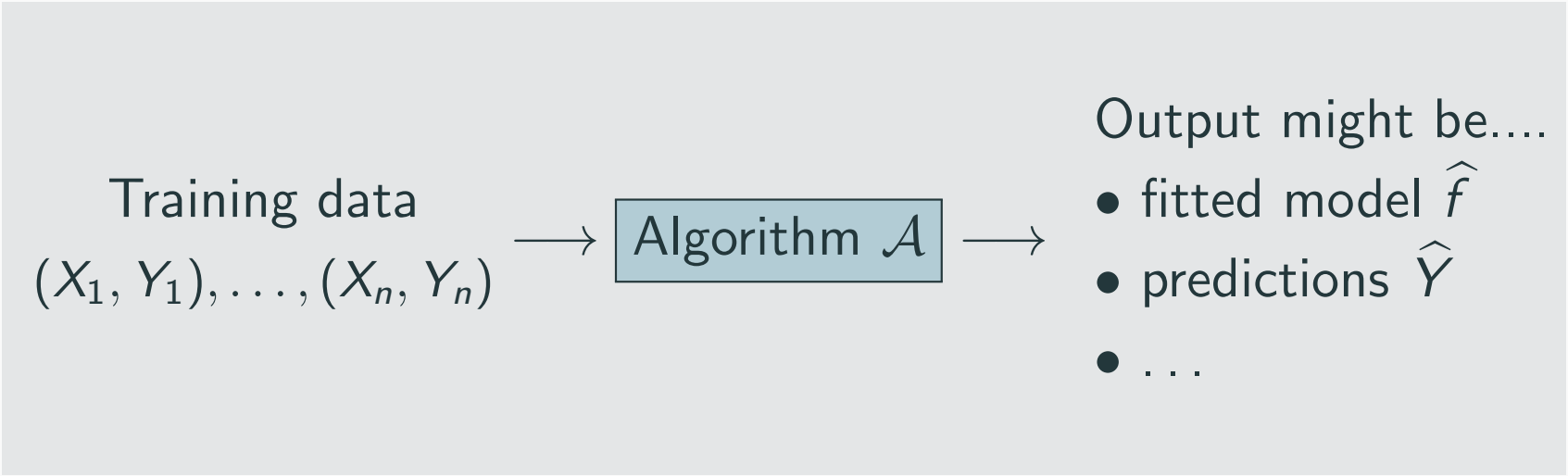
Müller, Luo, & B. 2025, *Are all models wrong? Fundamental limits in distribution-free empirical model falsification*

Background

Motivation


Training data
$(X_1, Y_1), \ldots, (X_n, Y_n)$
Algorithm $\mathcal{A}$
Output might be....
• fitted model $\widehat{f}$
• predictions $\widehat{Y}$
• ...

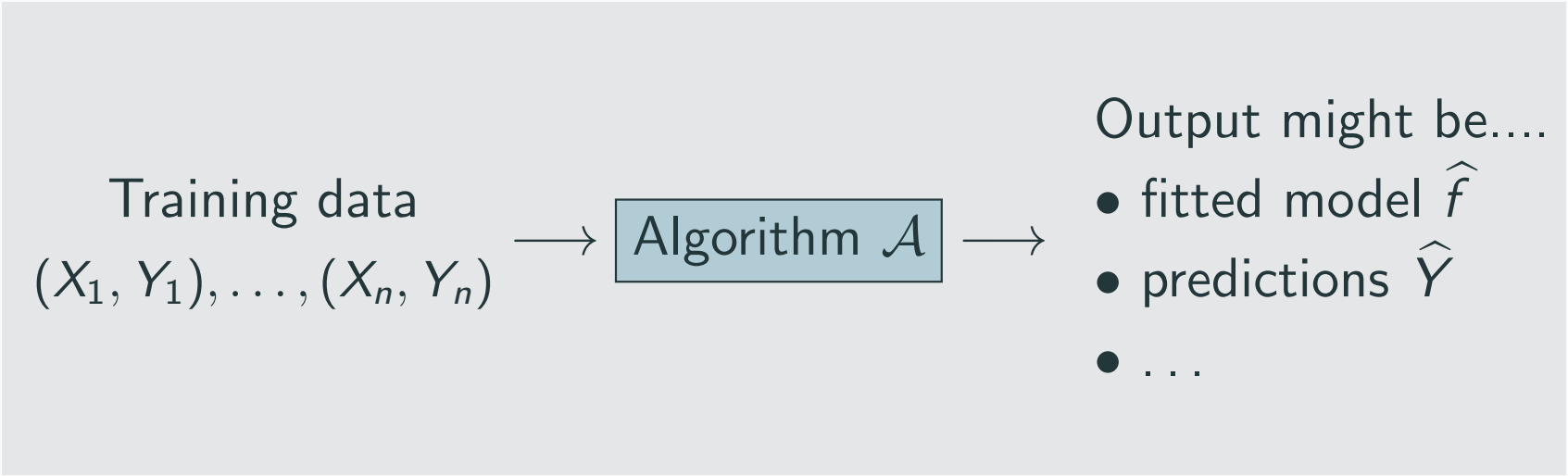
Motivation



The classical setting:

Strong assumptions $\Longrightarrow$ strong guarantees (e.g., consistency of $\widehat{f}$)

Motivation



The classical setting:

Strong assumptions $\Longrightarrow$ strong guarantees (e.g., consistency of $\widehat{f}$)

An assumption-lean setting:

Can we guarantee weaker properties (e.g., low prediction error), without strong assumptions?

Framework & notation

Defining the risk:

$$R_P(f) = \mathbb{E}_P\left[\ell(f(X), Y)\right]$$

loss function taking values in $[0, B]$
(e.g., squared error, 0/1 loss)

Framework & notation

Given an algorithm $\mathcal{A}$ that returns a fitted model...

- Can we evaluate the risk of models returned by $\mathcal{A}$?
- Can we compare the risk of models returned by $\mathcal{A}$ vs $\mathcal{A}'$?

Framework & notation

Given a model class $\mathcal{F}$ (e.g., all linear models)...

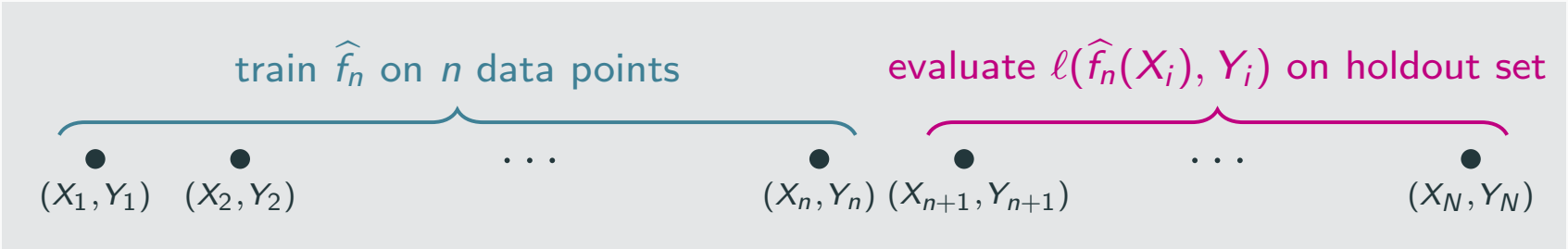
- Can we determine whether $\mathcal{F}$ is a good fit to the data—does any model $f \in \mathcal{F}$ have low risk?
- Can we test if $\mathcal{A}$ is finding a nearly-optimal model in $\mathcal{F}$?

Use a holdout set?

Assume we have access to a sample of size N:

$$(X_1, Y_1), \ldots, (X_N, Y_N) \overset{\text{iid}}{\sim} P$$

If we split data into a training set + a holdout set:



$\rightsquigarrow$ estimate $R_P(\widehat{f}_n)$ up to error $\asymp \frac{1}{\sqrt{N-n}}$

Use a holdout set?

If we use a holdout set to estimate $R_P(\widehat{f}_n)$ for fitted model $\widehat{f}_n$....

Can we determine whether $\mathcal{A}$ is a good algorithm?

- Risk of $\widehat{f}_n$ might be highly variable if resample training data

Can we determine if model class $\mathcal{F}$ is a good fit for the data?

- $\widehat{f}_n$ might be far from optimal within this model class

Use a holdout set?

If we use a holdout set to estimate $R_P(\widehat{f}_n)$ for fitted model $\widehat{f}_n$....

(I) Can we determine whether $\mathcal{A}$ is a good algorithm?

- Risk of $\widehat{f}_n$ might be highly variable if resample training data

(II) Can we determine if model class $\mathcal{F}$ is a good fit for the data?

- $\widehat{f}_n$ might be far from optimal within this model class

This talk: fundamental limits in the distribution-free regime.

(I) Hardness of testing algorithmic risk

Defining a black-box test

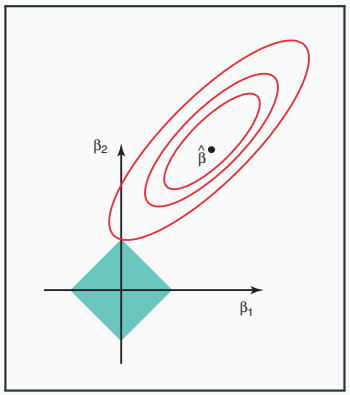
Given an algorithm $\mathcal{A}$, we might want to certify that $\mathcal{A}$ satisfies:

- Bounds on estimation error
- Bounds on prediction error or risk
- Properties like stability, privacy, robustness, interpretability, ...

The black-box setting: learn how $\mathcal{A}$ works by running it on data.

Defining a black-box test

Why black-box?



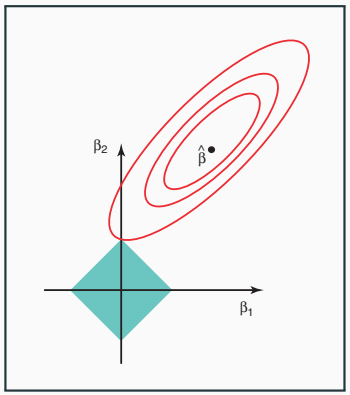
[1]

$\rightsquigarrow$ can study $\mathcal{A}$ theoretically & empirically

[1] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

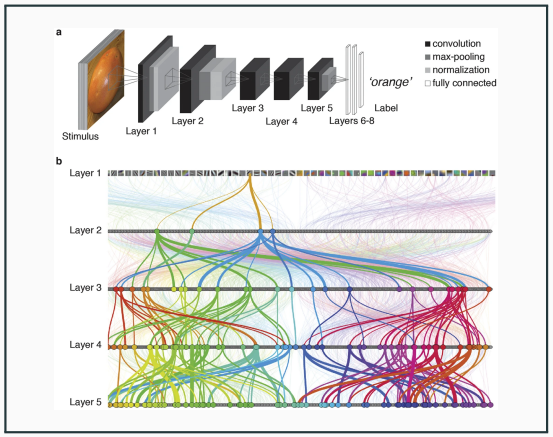
Defining a black-box test

Why black-box?



1

$\rightsquigarrow$ can study $\mathcal{A}$ theoretically & empirically



2

$\rightsquigarrow$ theoretical guarantees are challenging, but can study $\mathcal{A}$ empirically

[1] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

[2] Figure from Cichy et al 2016, *Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence*

Defining a black-box test

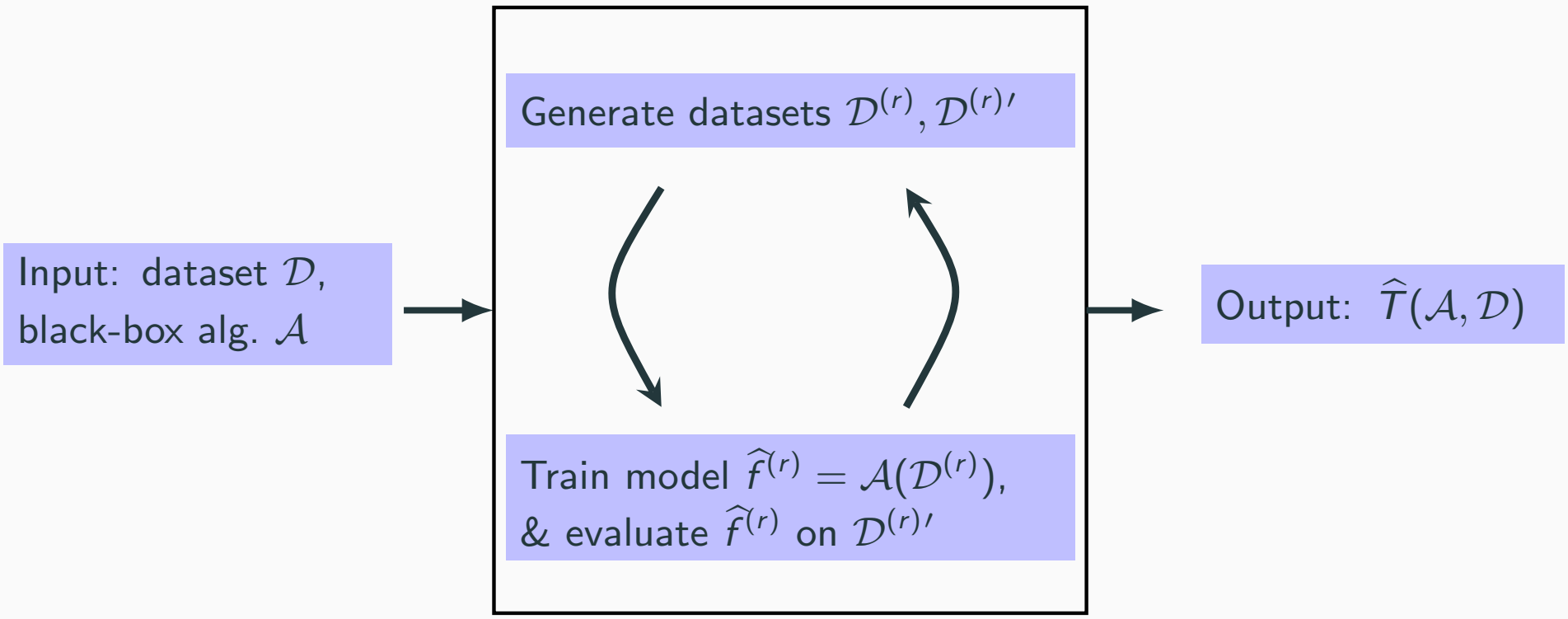
The black-box setting: learn how $\mathcal{A}$ works by running it on data.

For example, we may run $\mathcal{A}$ on....

- Subsets of the available real data
- Samples bootstrapped from available real data
- Semisynthetic data obtained by perturbing the real data
- Simulated data obtained by fitting a model to real data
- Etc.

Defining a black-box test

Iterate for rounds $r = 1, 2, \ldots$
(stop at some finite time)



(Can also incorporate randomization into $\mathcal{A}$ and/or into the test)

Framework

For evaluating $\mathcal{A}$—which question do we want to answer?[3]

model trained by $\mathcal{A}$ on dataset of size n

`EvaluateModel`: what is $R_P(\widehat{f}_n)$?

versus

`EvaluateAlg`: what is $R_{P,n}(\mathcal{A})$?

$\mathbb{E}_P\left[R_P(\widehat{f}_n)\right]$ for data $\overset{\text{iid}}{\sim} P$

[3] See discussion in:
Dietterich 1998, *Approximate statistical tests for comparing supervised classification learning algorithms*
Hastie, Tibshirani, Friedman 2009, *The elements of statistical learning: data mining, inference, and prediction*
Trippe, Deshpande, Broderick (2023), *Confidently comparing estimates with the c-value*

Framework

Given $N > n$ data points....

- Answering `EvaluateModel` is straightforward
  (use a holdout set of size $N - n$)

Framework

Given $N > n$ data points....

- Answering `EvaluateModel` is straightforward
  (use a holdout set of size $N - n$)

- Is it possible to answer `EvaluateAlg`?
  For instance can we perform a hypothesis test:

$$H_0 : \ R_{P,n}(\mathcal{A}) \geq \tau, \quad H_1 : \ R_{P,n}(\mathcal{A}) < \tau$$

Framework

A related question:

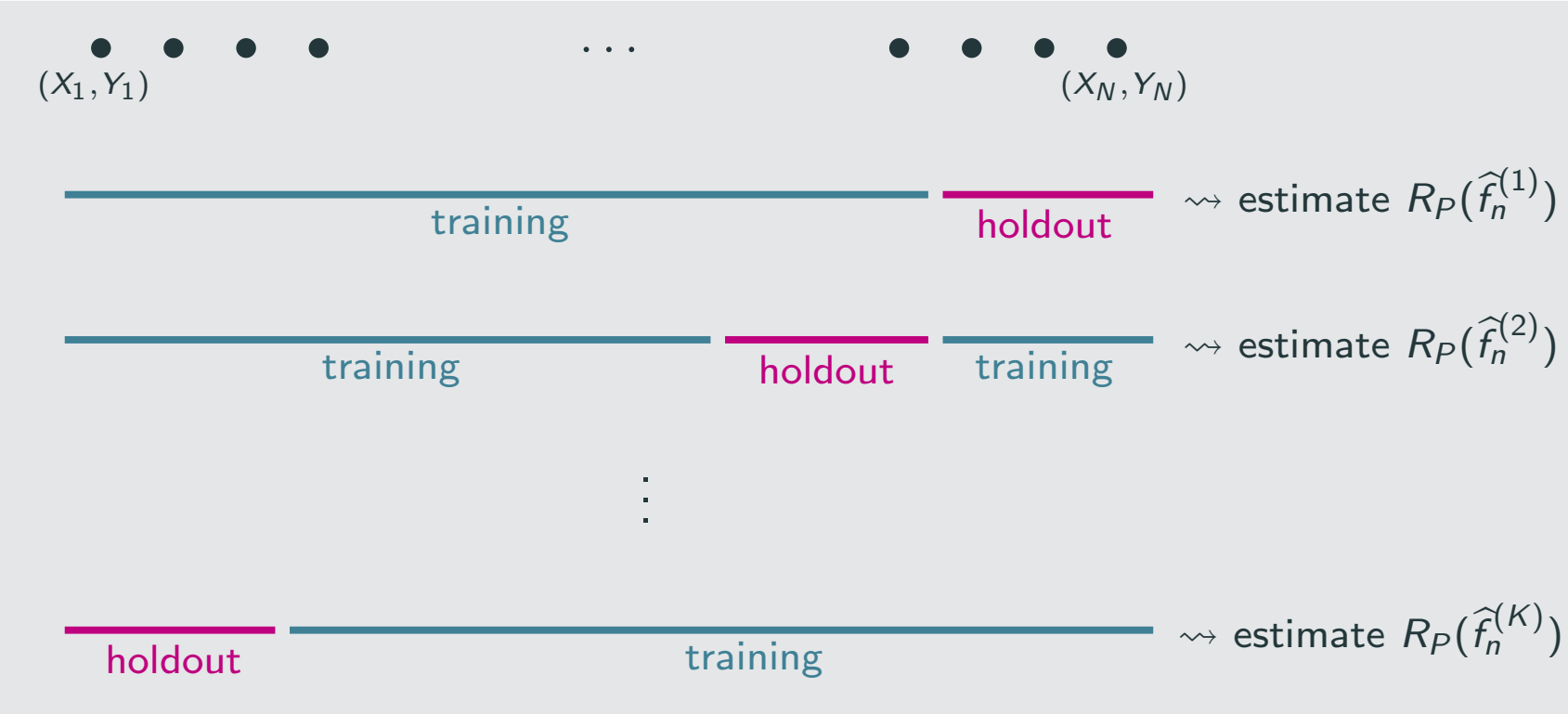
For comparing $\mathcal{A}$ & $\mathcal{A}'$—which question do we want to answer?

$$\texttt{CompareModel:}\ \text{is}\ R_P(\widehat{f}_n) < R_P(\widehat{f}'_n)?$$

versus

$$\texttt{CompareAlg:}\ \text{is}\ R_{P,n}(\mathcal{A}) < R_{P,n}(\mathcal{A}')?$$

Use cross-validation?



- For each fold $k = 1, \ldots, K$ we can estimate $R_P(\widehat{f}_n^{(k)})$
- Can we use the mean to estimate $R_{P,n}(\mathcal{A})$?[4]

[4] Bates, Hastie, Tibshirani, *Cross-validation: what does it estimate and how well does it do it?*

Defining Type I error

Hypothesis test:

- Testing $H_0: \ R_{P,n}(\mathcal{A}) \geq \tau$ vs. $H_1: \ R_{P,n}(\mathcal{A}) < \tau$
- Given a dataset $\mathcal{D}_N \overset{\text{iid}}{\sim} P$, return $\widehat{T}(\mathcal{A}, \mathcal{D}_N) \in \{0, 1\}$

Distribution-free validity

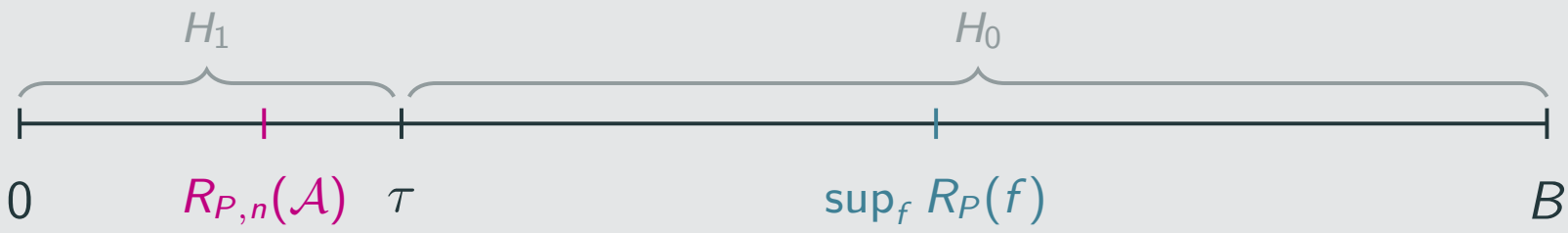
A test $\widehat{T}$ has distribution-free Type I error $\leq \alpha$ if

$$\underbrace{\mathbb{P}_P\left\{\widehat{T}(\mathcal{A}, \mathcal{D}_N) = 1\right\} \leq \alpha}_{\text{falsely claim that } R_{P,n}(\mathcal{A}) < \tau} \text{ for all } \mathcal{A}, P \text{ with } R_{P,n}(\mathcal{A}) \geq \tau$$

Defining signal strength

Defining signal strength

Defining signal strength



Defining signal strength

A measure of signal strength: $\dfrac{\tau - R_{P,n}(\mathcal{A})}{\sup_f R_P(f) - \tau}$

A hardness result

In the setting where $|\mathcal{X} \times \mathcal{Y}| = \infty \ldots$

Theorem: hardness of `EvaluateAlg`

For any black-box test $\widehat{T}$ with distribution-free Type I error $\leq \alpha$, for any $\mathcal{A}$ the power is bounded as

$$\mathbb{P}_P\left\{\widehat{T}(\mathcal{A}, \mathcal{D}_N) = 1\right\} \leq \alpha\left(1 + \frac{\overbrace{\tau - R_{P,n}(\mathcal{A})}^{\text{signal strength}}}{\sup_f R_P(f) - \tau} + \mathcal{O}(\tfrac{1}{N})\right)^{N/n}$$

- Similar results for `CompareAlg`
- See paper for a simple test achieving nearly optimal power

A hardness result

Interpretation:

- *Every* valid test has low power: if $N = \mathcal{O}(n)$ and τ small, then power $\lessapprox \alpha$ (no better than random)
- Cross-validation does not give assumption-free guarantees for estimating/testing $R_{P,n}(\mathcal{A})$

Proof sketch for hardness result

- Choose a function f_* such that $R_P(f_*) \approx \sup_f R_P(f)$
- Choose some $(x_*, y_*) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{P}_P\Big\{ \underbrace{(x_*, y_*) \in \mathcal{D}_N \cup \mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \dots}_{\text{the data point } (x_*, y_*) \text{ occurs when running } \widehat{T}} \Big\} \approx 0$$

Proof sketch for hardness result

- Choose a function f_* such that $R_P(f_*) \approx \sup_f R_P(f)$
- Choose some $(x_*, y_*) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{P}_P\Big\{ \underbrace{(x_*, y_*) \in \mathcal{D}_N \cup \mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \dots}_{\text{the data point } (x_*, y_*) \text{ occurs when running } \widehat{T}} \Big\} \approx 0$$

The construction

Given any algorithm $\mathcal{A}$ and distribution P, define

$$P' = (1-c) \cdot P + c \cdot \delta_{(x_*, y_*)}, \quad \mathcal{A}'(\mathcal{D}) = \begin{cases} \mathcal{A}(\mathcal{D}), & \text{if } (x_*, y_*) \notin \mathcal{D} \\ f_*, & \text{if } (x_*, y_*) \in \mathcal{D} \end{cases}$$

Proof sketch for hardness result

- Choose a function f_* such that $R_P(f_*) \approx \sup_f R_P(f)$
- Choose some $(x_*, y_*) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{P}_P\Big\{ \underbrace{(x_*, y_*) \in \mathcal{D}_N \cup \mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \dots}_{\text{the data point } (x_*, y_*) \text{ occurs when running } \widehat{T}} \Big\} \approx 0$$

The construction

Given any algorithm $\mathcal{A}$ and distribution P, define

$$P' = (1-c) \cdot P + c \cdot \delta_{(x_*, y_*)}, \quad \mathcal{A}'(\mathcal{D}) = \begin{cases} \mathcal{A}(\mathcal{D}), & \text{if } (x_*, y_*) \notin \mathcal{D} \\ f_*, & \text{if } (x_*, y_*) \in \mathcal{D} \end{cases}$$

Choose c so that $R_{P',n}(\mathcal{A}') \geq \tau$,

but $(\mathcal{A}, P)$ and $(\mathcal{A}', P')$ are not easily distinguishable

Connecting risk & stability

Recall the questions:

`EvaluateModel`: what is $R_P(\widehat{f}_n)$?

versus

`EvaluateAlg`: what is $R_{P,n}(\mathcal{A})$?

[5] Dietterich 1998, *Approximate statistical tests for comparing supervised classification learning algorithms*

Connecting risk & stability

Recall the questions:

`EvaluateModel`: what is $R_P(\widehat{f}_n)$?

versus

`EvaluateAlg`: what is $R_{P,n}(\mathcal{A})$?

- Testing `EvaluateAlg` is harder due to variability in $\widehat{f}_n$
- If assume $\mathcal{A}$ is stable, does `EvaluateAlg` become testable?[5]

[5] Dietterich 1998, *Approximate statistical tests for comparing supervised classification learning algorithms*

Stability or consistency?

An ℓ_2 definition of stability

$\mathcal{A}$ is β^2-stable w.r.t. distribution P and sample size n if

$$\mathbb{E}_P\left[\left(\widehat{f}_n(X_{n+1}) - \widehat{f}_{n-1}(X_{n+1})\right)^2\right] \leq \beta^2$$

Stability or consistency?

An ℓ_2 definition of stability

$\mathcal{A}$ is β^2-stable w.r.t. distribution P and sample size n if

$$\mathbb{E}_P\left[\left(\widehat{f}_n(X_{n+1}) - \widehat{f}_{n-1}(X_{n+1})\right)^2\right] \leq \beta^2$$

By the Efron–Stein inequality,

$$\mathbb{E}_P\left[\left(\widehat{f}_n(X_{n+1}) - \bar{f}(X_{n+1})\right)^2\right] \leq n \cdot \beta^2$$

where $\bar{f}(x) := \mathbb{E}_P\left[\widehat{f}_n(x)\right]$

and therefore,

$$\beta^2\text{-stability with } \beta = o(n^{-1/2}) \implies \text{consistency, i.e., } \widehat{f}_n \approx \bar{f}$$

Stability or consistency?

Stability of the loss

$\mathcal{A}$ is β^2-loss-stable w.r.t. distribution P and sample size n if

$$\mathbb{E}_P\left[\left(\ell\big(\widehat{f}_n(X_{n+1}), Y_{n+1}\big) - \ell\big(\widehat{f}_{n-1}(X_{n+1}), Y_{n+1}\big)\right)^2\right] \leq \beta^2$$

Stability or consistency?

Stability of the loss

$\mathcal{A}$ is β^2-loss-stable w.r.t. distribution P and sample size n if

$$\mathbb{E}_P\left[\left(\ell\big(\widehat{f}_n(X_{n+1}), Y_{n+1}\big) - \ell\big(\widehat{f}_{n-1}(X_{n+1}), Y_{n+1}\big)\right)^2\right] \leq \beta^2$$

By the Efron–Stein inequality,

$$\mathbb{E}_P\left[\big(R_P(\widehat{f}_n) - R_{P,n}(\mathcal{A})\big)^2\right] \leq n \cdot \beta^2$$

$\rightsquigarrow$ β^2-loss-stability with $\beta = o(n^{-1/2})$ implies[6]

$$\texttt{EvaluateModel} \approx \texttt{EvaluateAlg}$$

[6] Results on cross-validation based estimation of $R_{P,n}(\mathcal{A})$ in this setting:
Austern et al 2020, *Asymptotics of cross-validation*
Bayle et al 2020, *Cross-validation confidence intervals for test error*

Stability or consistency?

If instead $\beta \geq 2Bn^{-1/2}$, then get same hardness result as before:

Theorem: hardness of `EvaluateAlg` with stability

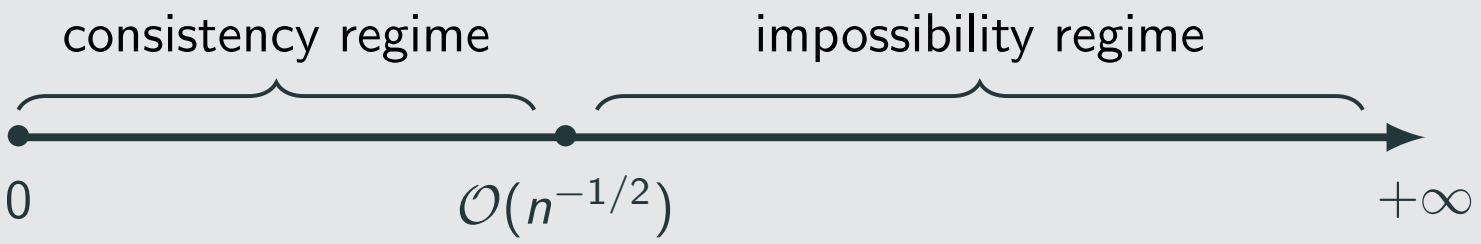
Let $\widehat{T}$ be any black-box test,
with Type I error $\leq \alpha$ for all β^2-loss-stable $\mathcal{A}, P$.

Then for any $\mathcal{A}$ the power is bounded as

$$\mathbb{P}_P\left\{\widehat{T}(\mathcal{A}, \mathcal{D}_N) = 1\right\} \leq \alpha\left(1 + \frac{\overbrace{\tau - R_{P,n}(\mathcal{A})}^{\text{signal strength}}}{\sup_f R_P(f) - \tau} + \mathcal{O}(\tfrac{1}{N})\right)^{N/n}$$

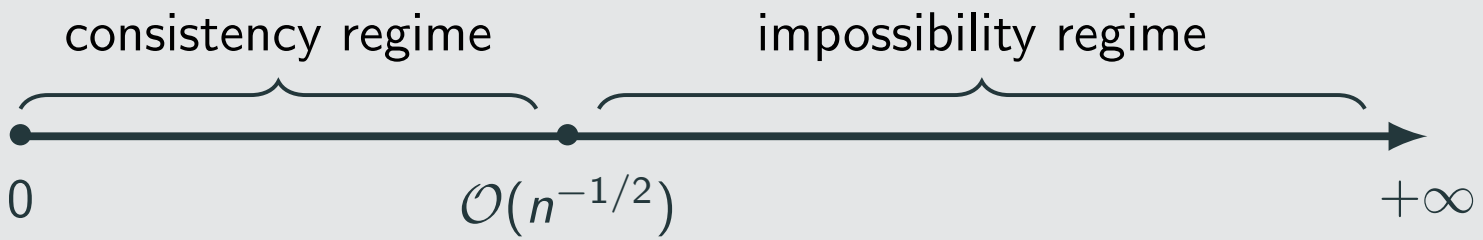
Stability or consistency?

Overview:



Stability or consistency?

Overview:



Interpretation:

- Even if assume stability, impossible to answer `EvaluateAlg`...
- ....unless we assume such strong stability that $\widehat{f}_n$ concentrates

(II) **Testing a model class**

Framework

Goal: given a model class $\mathcal{F}$, perform inference on

$$R_P(\mathcal{F}) = \inf_{f \in \mathcal{F}} R_P(f)$$

using data sampled from P.

Framework

Goal: given a model class $\mathcal{F}$, perform inference on

$$R_P(\mathcal{F}) = \inf_{f \in \mathcal{F}} R_P(f)$$

using data sampled from P.

- Supervised learning setting: data $(X_i, Y_i) \overset{\text{iid}}{\sim} P$, and

$$R_P(f) = \mathbb{E}_P\left[\ell(f(X), Y)\right]$$

- More generally: data $Z_i \overset{\text{iid}}{\sim} P$, and

$$R_P(f) = \mathbb{E}_P\left[\ell(f, Z)\right]$$

Framework

Motivation:

- Goodness-of-fit: is $\mathcal{F}$ a good choice for modeling the data?
- Signal-to-noise ratio: how "noisy" is Y conditional on X?
- Evaluating an algorithm: if $\mathcal{A}$ returns a fitted model $\widehat{f}_n \in \mathcal{F}$, can we perform inference on its *excess risk*?

$$\text{ExcessRisk}(\hat{f}_n) = \underbrace{R_P(\hat{f}_n)}_{\texttt{EvaluateModel}\ \checkmark} - \underbrace{\inf_{f \in \mathcal{F}} R_P(f)}_{???}$$

Upper & lower bounds

Given a dataset $\mathcal{D}_n \overset{\text{iid}}{\sim} P$ and a model class $\mathcal{F}$,

can we provide upper & lower bounds on $R_P(\mathcal{F})$?

Distribution-free validity for a lower bound

For all distributions P on $\mathcal{Z}$,

$$\mathbb{P}_P \left\{ R_P(\mathcal{F}) \geq \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) \right\} \geq 1 - \alpha$$

Distribution-free validity for an upper bound

For all distributions P on $\mathcal{Z}$,

$$\mathbb{P}_P \left\{ R_P(\mathcal{F}) \leq \hat{U}_\alpha(\mathcal{F}, \mathcal{D}_n) \right\} \geq 1 - \alpha$$

Upper & lower bounds

An asymmetry in the problem:

- Upper bounds are easier: given fitted model $\widehat{f}_n$,

$$R_P(\mathcal{F}) = \inf_{f \in \mathcal{F}} R_P(f) \leq \underbrace{R_P(\widehat{f}_n)}_{\text{can use holdout set to construct valid upper bound}}$$

- Lower bounds are harder: if $R_P(\widehat{f}_n)$ is large, is this because....
  - $R_P(\mathcal{F})$ is large (i.e., $\mathcal{F}$ is not a good fit to the data)?
  - Or, $\widehat{f}_n$ is far from optimal (i.e., $\mathcal{A}$ is not a good alg.)?

Upper & lower bounds

A trivial solution:

$$\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = \begin{cases} 0, & \text{with probability } 1-\alpha \\ +\infty, & \text{with probability } \alpha \end{cases}$$

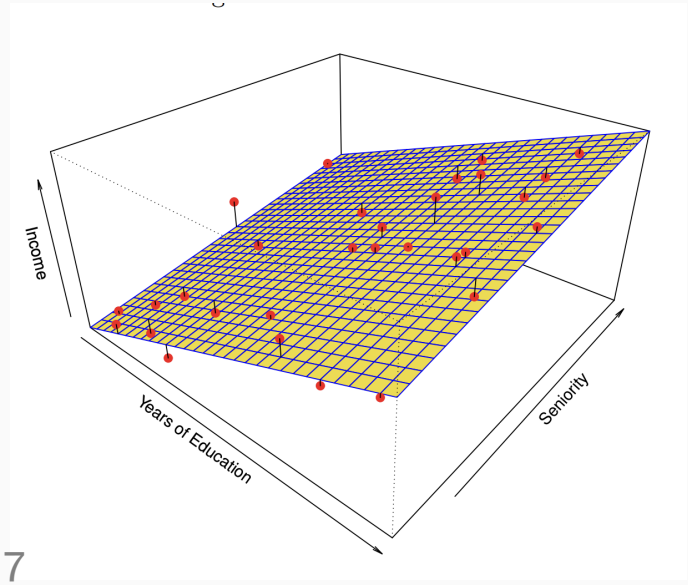
Upper & lower bounds

A trivial solution:

$$\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = \begin{cases} 0, & \text{with probability } 1-\alpha \\ +\infty, & \text{with probability } \alpha \end{cases}$$

$\rightsquigarrow$ any *nontrivial* solution must have $\mathbb{P}_P\left\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\right\} > \alpha$

Connections to generalization & interpolation



[7]

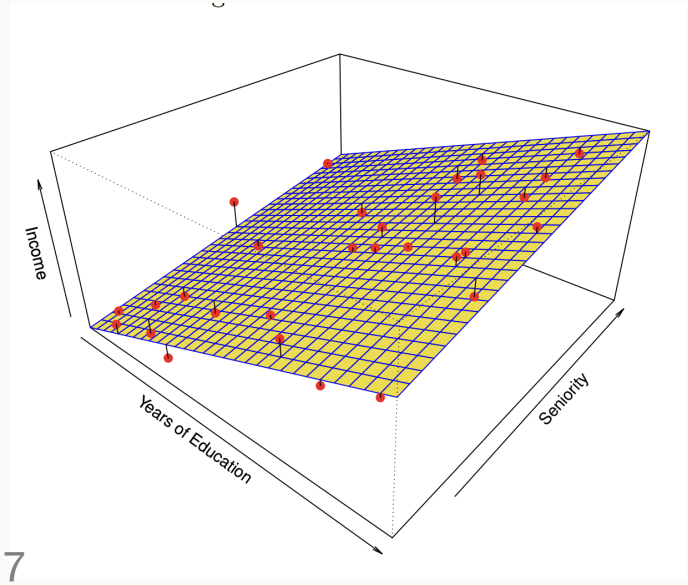
If $\mathcal{F}$ is low-dim., can use generalization:

$$\mathbb{P}_P \left\{ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R_P(f)| \leq \epsilon_n \right\} \geq 1 - \alpha$$

empirical risk $\frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i)$

[7] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

Connections to generalization & interpolation



7

If $\mathcal{F}$ is low-dim., can use generalization:

$$\mathbb{P}_P\left\{\sup_{f\in\mathcal{F}}|\hat{R}_n(f)-R_P(f)|\le\epsilon_n\right\}\ge 1-\alpha$$

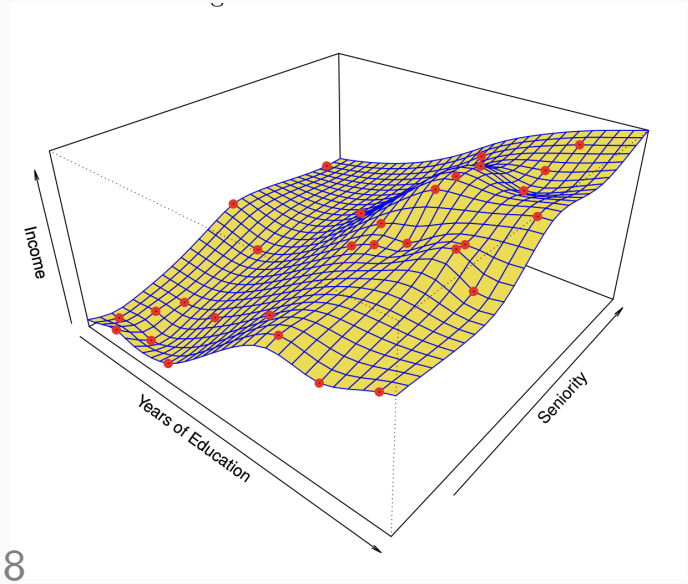
empirical risk $\frac{1}{n}\sum_{i=1}^n \ell(f,Z_i)$

Valid distribution-free bounds for $R_P(\mathcal{F})$:

$$\hat{L}_\alpha(\mathcal{F},\mathcal{D}_n)=\inf_{f\in\mathcal{F}}\hat{R}_n(f)-\epsilon_n,\quad \hat{U}_\alpha(\mathcal{F},\mathcal{D}_n)=\inf_{f\in\mathcal{F}}\hat{R}_n(f)+\epsilon_n$$

[7] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

Connections to generalization & interpolation



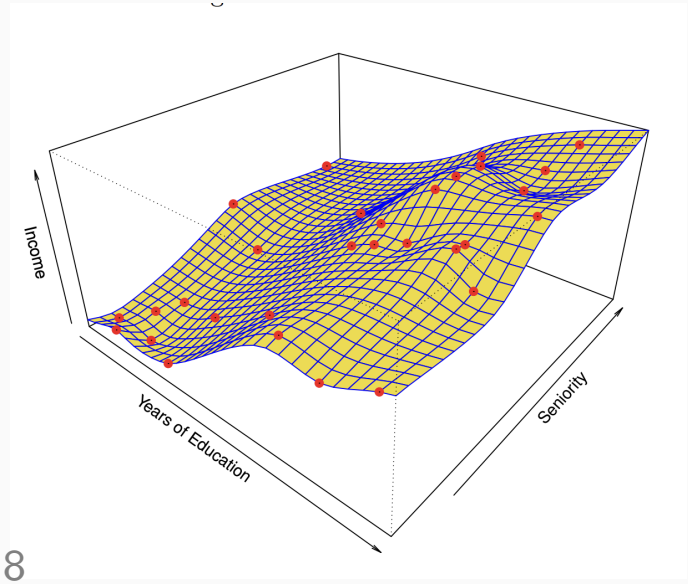
[8]

If $\mathcal{F}$ has high complexity,

some $f \in \mathcal{F}$ will interpolate the data

[8] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

Connections to generalization & interpolation



[8]

If $\mathcal{F}$ has high complexity,

some $f \in \mathcal{F}$ will interpolate the data

If we observe interpolation, i.e., $\inf_{f \in \mathcal{F}} \hat{R}_n(f) = 0$....

- Can we ever be *confident* that interpolation is solely due to complexity of $\mathcal{F}$, and in fact $R_P(\mathcal{F})$ is large?
- Or, is $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = 0$ the only valid lower bound?

[8] Figure from James, Witten, Hastie, Tibshirani 2013, *An Introduction to Statistical Learning*

A valid lower bound via ERM

Empirical risk minimization: $\hat{R}_n(\mathcal{F}) = \inf_{f \in \mathcal{F}} \hat{R}_n(f)$

Theorem

If Δ_n is the unique solution to $-\Delta_n - \log(1 - \Delta_n) = \frac{B \log(1/\alpha)}{n \hat{R}_n(\mathcal{F})}$, then

$$\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) = (1 - \Delta_n) \cdot \hat{R}_n(\mathcal{F})$$

is a valid distribution-free lower bound.

A valid lower bound via ERM

Interpreting this solution...

- If $\hat{R}_n(\mathcal{F}) > 0$ then this is a nontrivial lower bound:

$$\hat{R}_n(\mathcal{F}) > 0 \;\Rightarrow\; \hat{L}_\alpha^{\mathsf{ERM}}(\mathcal{F}, \mathcal{D}_n) > 0$$

A valid lower bound via ERM

Interpreting this solution...

- If $\hat{R}_n(\mathcal{F}) > 0$ then this is a nontrivial lower bound:

$$\hat{R}_n(\mathcal{F}) > 0 \;\Rightarrow\; \hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) > 0$$

$$\geq \hat{R}_n(\mathcal{F}) - \mathcal{O}(n^{-1/2})$$

A valid lower bound via ERM

Interpreting this solution...

- If $\hat{R}_n(\mathcal{F}) > 0$ then this is a nontrivial lower bound:

$$\hat{R}_n(\mathcal{F}) > 0 \;\Rightarrow\; \hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) > 0$$

$\nwarrow$ $\geq \hat{R}_n(\mathcal{F}) - \mathcal{O}(n^{-1/2})$

- But under interpolation....

$$\hat{R}_n(\mathcal{F}) = 0 \;\Rightarrow\; \hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) = 0$$

Is *any* valid & nontrivial lower bound possible, in this regime?

Interpolation capacity

We say that $\mathcal{F}$ *interpolates* a dataset $\mathcal{D}_n$ if $\hat{R}_n(\mathcal{F}) = 0$.

Interpolation capacity of $\mathcal{F}$

$$N(\mathcal{F}, P) = \sup\left\{n : \mathbb{P}_P\left\{\mathcal{F} \text{ interpolates } \mathcal{D}_n\right\} = 1\right\}$$

$$N_+(\mathcal{F}, P) = \sup\left\{n : \mathbb{P}_P\left\{\mathcal{F} \text{ interpolates } \mathcal{D}_n\right\} > 0\right\}$$

Interpolation capacity

We say that $\mathcal{F}$ *interpolates* a dataset $\mathcal{D}_n$ if $\hat{R}_n(\mathcal{F}) = 0$.

Interpolation capacity of $\mathcal{F}$

$$N(\mathcal{F}, P) = \sup\{n : \mathbb{P}_P\{\mathcal{F} \text{ interpolates } \mathcal{D}_n\} = 1\}$$

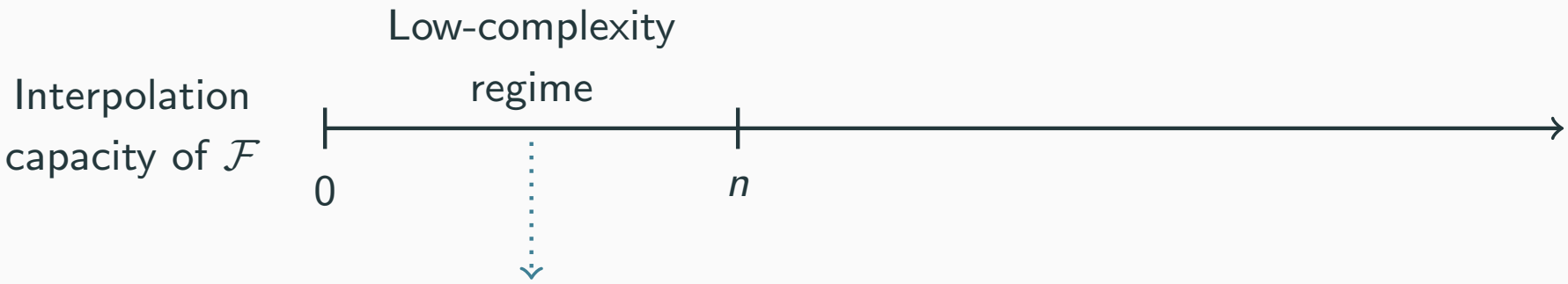
$$N_+(\mathcal{F}, P) = \sup\{n : \mathbb{P}_P\{\mathcal{F} \text{ interpolates } \mathcal{D}_n\} > 0\}$$

Example:

- $\mathcal{F} = \{\text{linear models in dim. } d\}$
  Then $N(\mathcal{F}, P) = N_+(\mathcal{F}, P) = d$ for any continuous distrib. P

Overview of results (so far)

Overview of results (so far)



A valid & nontrivial lower bound exists (via the ERM)

A hardness result

- If $\hat{R}_n(\mathcal{F}) > 0$ then the ERM provides a positive lower bound
- What happens if $\hat{R}_n(\mathcal{F}) = 0$ (i.e., $\mathcal{F}$ interpolates $\mathcal{D}_n$)?

A hardness result

- If $\hat{R}_n(\mathcal{F}) > 0$ then the ERM provides a positive lower bound
- What happens if $\hat{R}_n(\mathcal{F}) = 0$ (i.e., $\mathcal{F}$ interpolates $\mathcal{D}_n$)?

Theorem

For any valid distribution-free lower bound,

$$\mathbb{P}_P \left\{ \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0 \right\} \leq \alpha + \frac{n^2}{2N(\mathcal{F}, P)}$$

Interpretation: if $N(\mathcal{F}, P) \gg n^2$ then any valid lower bound is essentially trivial.

Proof sketch

Let $Z_1, \ldots, Z_{N(\mathcal{F},P)} \overset{\text{iid}}{\sim} P$, and let Q be the empirical distribution.

By def. of interpolation capacity, almost surely,

$$\mathcal{F} \text{ interpolates this dataset} \rightsquigarrow R_Q(\mathcal{F}) = 0$$

Proof sketch

Let $Z_1, \ldots, Z_{N(\mathcal{F},P)} \overset{\text{iid}}{\sim} P$, and let Q be the empirical distribution.

By def. of interpolation capacity, almost surely,

$$\mathcal{F} \text{ interpolates this dataset} \rightsquigarrow R_Q(\mathcal{F}) = 0$$

By validity of the lower bound, $\mathbb{P}_Q \left\{ \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0 \right\} \leq \alpha$

Proof sketch

Let $Z_1, \ldots, Z_{N(\mathcal{F},P)} \overset{\text{iid}}{\sim} P$, and let Q be the empirical distribution.

By def. of interpolation capacity, almost surely,

$$\mathcal{F} \text{ interpolates this dataset} \rightsquigarrow R_Q(\mathcal{F}) = 0$$

By validity of the lower bound, $\mathbb{P}_Q \left\{ \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0 \right\} \leq \alpha$

Finally, the following distributions have TV distance $\leq \frac{n^2}{2N(\mathcal{F},P)}$:

- Sample n data points i.i.d. from P
- Construct a random Q, then sample n data points i.i.d. from Q

Proof sketch

Let $Z_1, \ldots, Z_{N(\mathcal{F},P)} \overset{\text{iid}}{\sim} P$, and let Q be the empirical distribution.

By def. of interpolation capacity, almost surely,

$$\mathcal{F} \text{ interpolates this dataset} \rightsquigarrow R_Q(\mathcal{F}) = 0$$

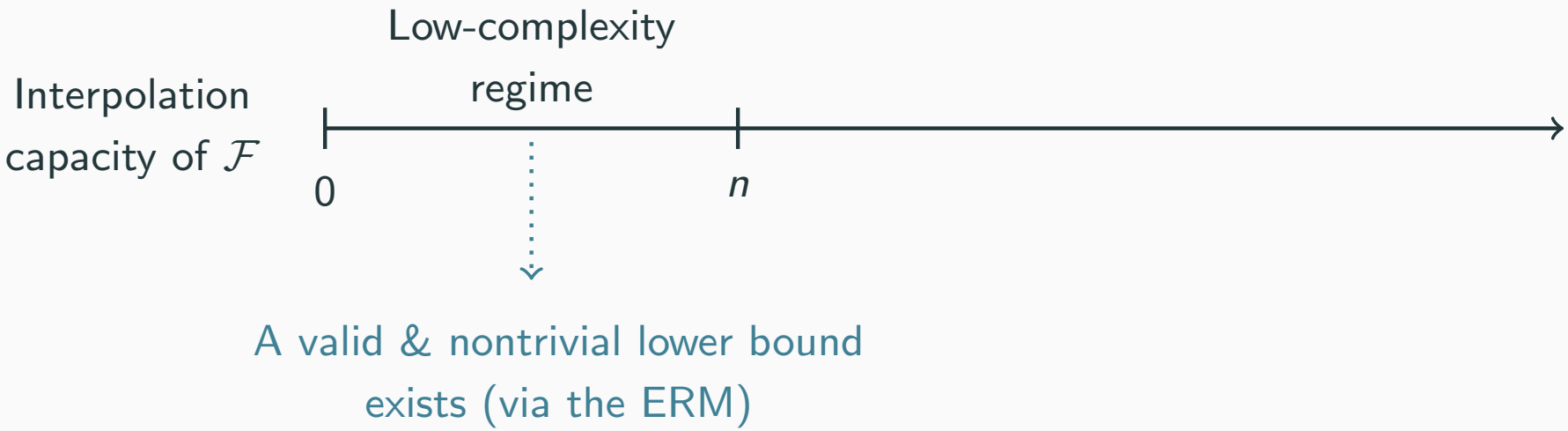
By validity of the lower bound, $\mathbb{P}_Q \left\{ \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0 \right\} \leq \alpha$

Finally, the following distributions have TV distance $\leq \frac{n^2}{2N(\mathcal{F},P)}$:

- Sample n data points i.i.d. from P
- Construct a random Q, then sample n data points i.i.d. from Q

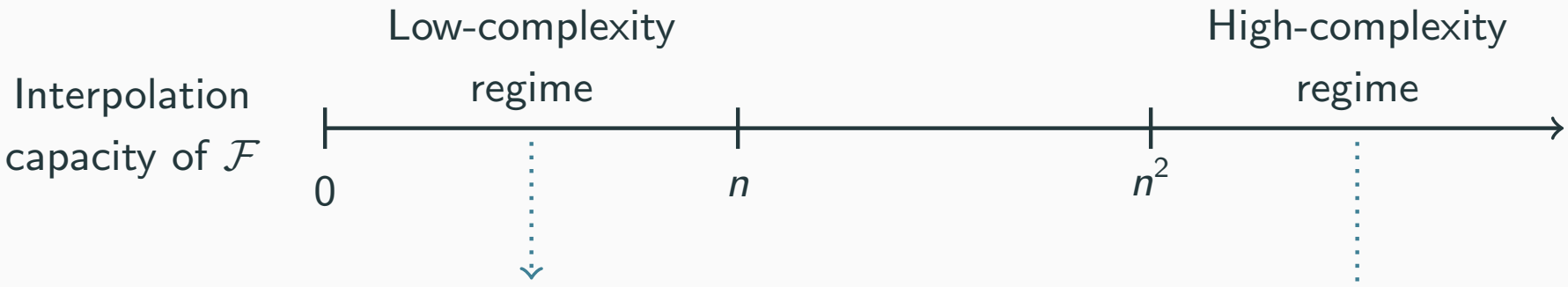
$$\implies \quad \mathbb{P}_P \left\{ \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0 \right\} \leq \alpha + \frac{n^2}{2N(\mathcal{F}, P)}$$

Overview of results (so far)



A valid & nontrivial lower bound exists (via the ERM)

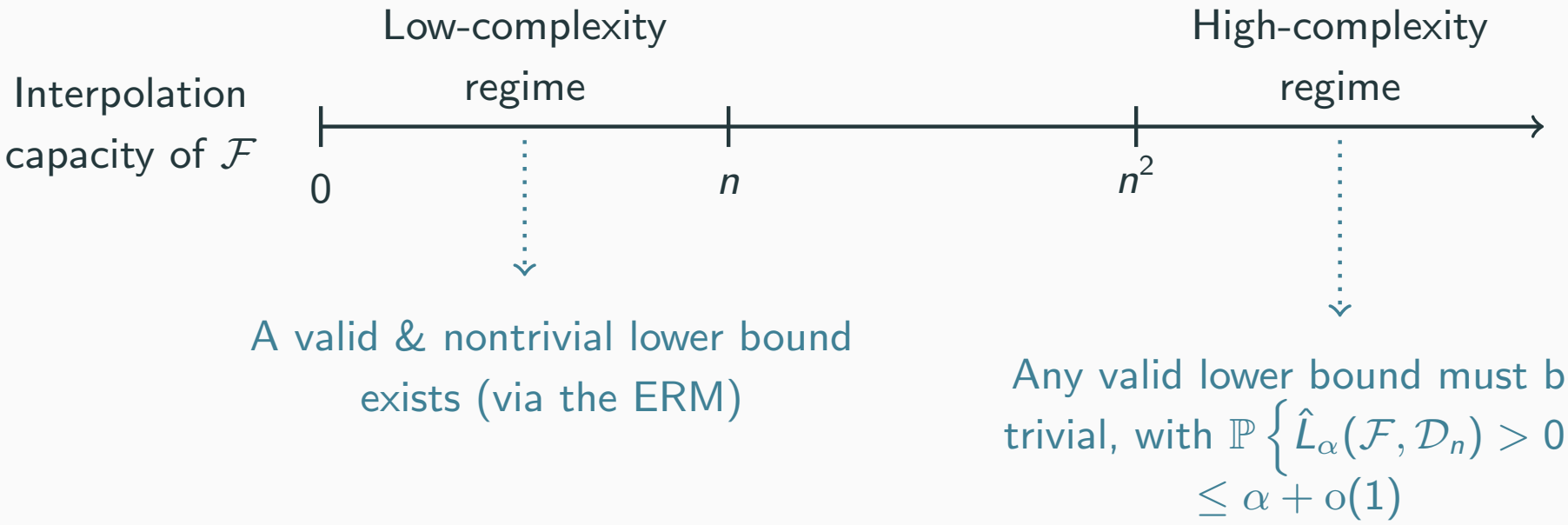
Overview of results (so far)



A valid & nontrivial lower bound exists (via the ERM)

Any valid lower bound must be trivial, with $\mathbb{P}\left\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\right\} \leq \alpha + o(1)$

Overview of results (so far)



Remaining question: what happens in between n and n^2?

(Is one of our bounds loose—is there a single phase transition?)

Example 1: linear models

$\mathcal{F}_{\text{lin}}^{(d)} = \{\text{all linear models on } \mathbb{R}^d\}, \quad P = \text{continuous distrib.}$

- Interpolation capacity $N(\mathcal{F}_{\text{lin}}^{(d)}, P) = d$

Example 1: linear models

$\mathcal{F}_{\text{lin}}^{(d)} = \{\text{all linear models on } \mathbb{R}^d\}, \quad P = \text{continuous distrib.}$

- Interpolation capacity $N(\mathcal{F}_{\text{lin}}^{(d)}, P) = d$

Theorem

For any valid distribution-free lower bound,

$$\mathbb{P}_P \left\{ \hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n) > 0 \right\} \leq \alpha + \frac{1}{2}\sqrt{\frac{n}{d-n-1}}$$

when P is multivariate Gaussian.

Interpretation: if $d \gg n$ then valid & nontrivial inference impossible (see paper for results on more general P)

Example 2: piecewise constant models

$\mathcal{F}_{\text{pwc}}^{(m)} = \{\text{functions taking} \leq m \text{ values}\}, \quad P = \text{continuous distrib.}$

- Interpolation capacity $N(\mathcal{F}_{\text{pwc}}^{(m)}, P) = m$

Example 2: piecewise constant models

$\mathcal{F}_{\text{pwc}}^{(m)} = \{\text{functions taking} \leq m \text{ values}\}$, $P = \text{continuous distrib.}$

- Interpolation capacity $N(\mathcal{F}_{\text{pwc}}^{(m)}, P) = m$

Theorem

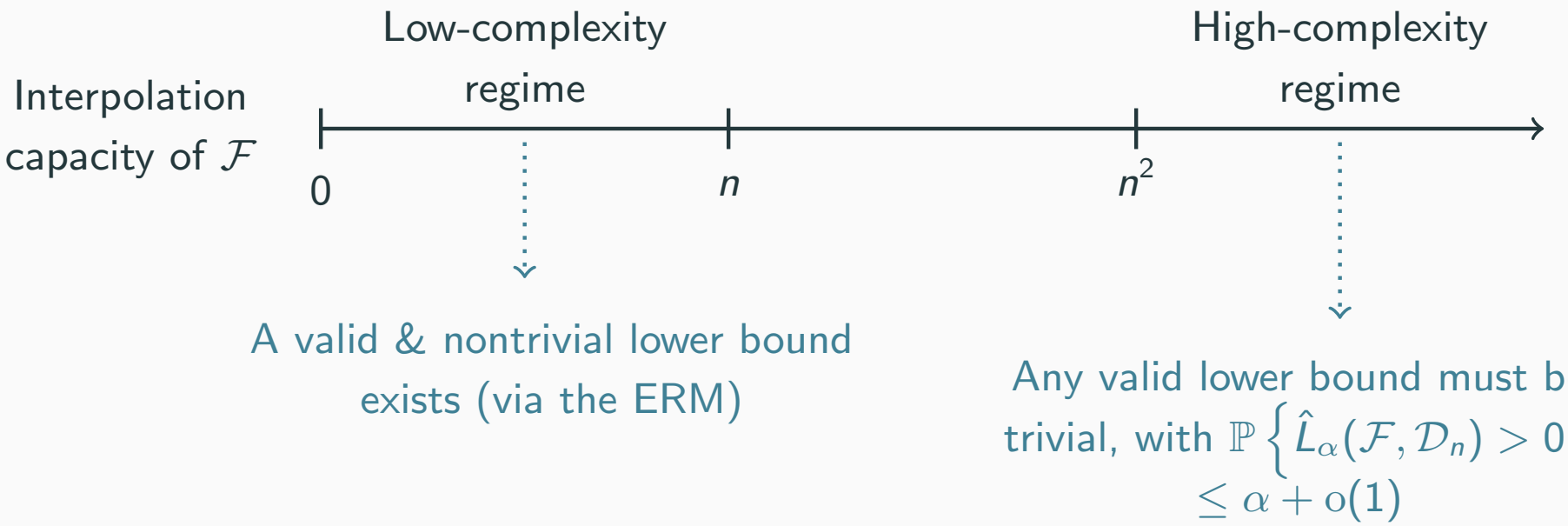
If $m < \frac{n(n-1)}{2\log(1/\alpha)}$, then there exists a valid distribution-free $\hat{L}_\alpha$ with

$$\mathbb{P}_P\left\{\hat{L}_\alpha(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) > 0\right\} = 1$$

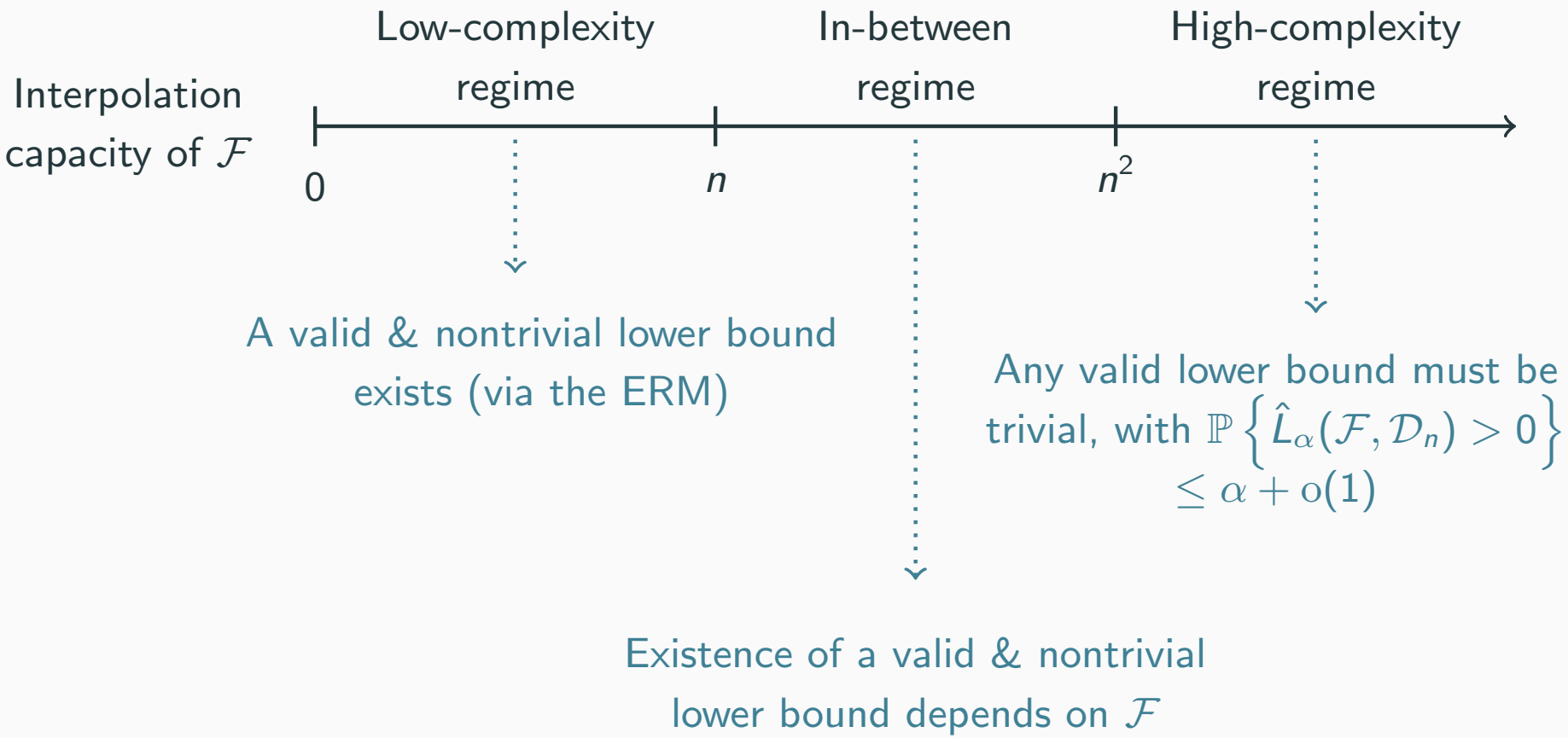
for every continuous distribution P

Interpretation: if $m \ll n^2$ then valid & nontrivial inference possible

Overview of results



Overview of results



Summary

Summary & open questions

Part (I) — inference on the risk of a black-box algorithm $\mathcal{A}$

If we treat any algorithm as a "black box"...

- Testing the risk of $\mathcal{A}$ is hard—all distrib.-free tests are essentially powerless (even w/ a weak stability assumption!)

Open questions: does testing risk become possible if we...

- Restrict to certain classes of algorithms?
- Add a post-processing step to our algorithms (e.g., bagging)?
- Use a different definition of risk?

Summary & open questions

Part (II) — inference on the risk of a model class $\mathcal{F}$

With no assumptions on the distribution P...

- The ERM enables inference on $R_P(\mathcal{F})$
  for low-dim. $\mathcal{F}$ (interpolation capacity $< n$)
- But, inference on $R_P(\mathcal{F})$ is hard
  for ultra-high-dim. $\mathcal{F}$ (interpolation capacity $\gg n^2$)

Open questions:

- Can we define a different notion of complexity of $\mathcal{F}$
  to eliminate the "in between" regime?
- Weakest possible assumptions on P to enable inference?