(Machine) Learning Parameter Regions

José Luis Montiel-Olea (Columbia)
James Nesbit (NYU)

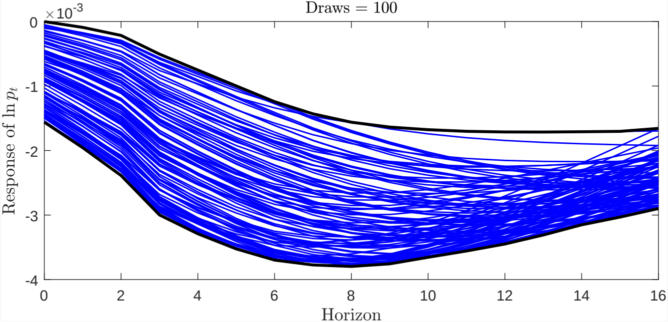
June 21st, 2018

Introduction

- Recent gainful connection between Machine Learning & Econometrics
    - "Machine learning is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction (regression), classification, and clustering or grouping tasks" (Athey 2018)
- We use off-the-shelf ML concepts to study computational issues arising in some econometric models
- Motivating example is SVARs, but the scope is more general
    - reporting an estimator of an identified set
    - confidence set formed via inverting test statistics
    - highest posterior density credible set for a vector-valued parameter

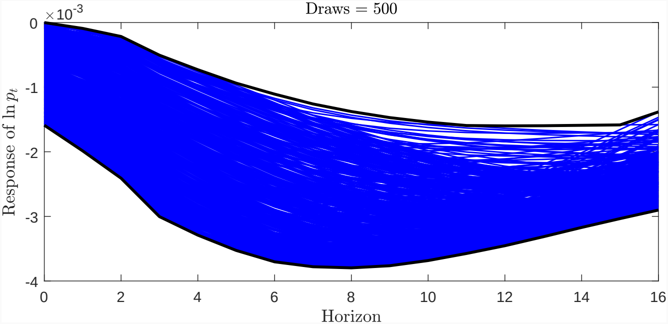
A Motivating Example

- How does the price level respond to a monetary policy contraction?
- Model + US data + theory restrict responses to a set
- Difficult to describe this set analytically, so use ‘random sampling’ to generate an approximation



A Motivating Example

- How does the price level respond to a monetary policy contraction?
- Model + US data + theory restrict responses to a set
- Difficult to describe this set analytically, so use 'random sampling' to generate an approximation



This Paper

- How many draws are needed to ‘learn’ these types of ‘parameter regions’?

This Paper

- How many draws are needed to ‘learn’ these types of ‘parameter regions’?
- ‘Random sampling approximations’ are a ‘supervised machine learning’ problems
    - Analogous to sampling pixels of an image to recognize it
- This analogy allows us to
    1. Build a framework of ‘learning’ to discipline the way we think about the accuracy of a random sampling approximation
    2. Characterize what can and cannot be learned
    3. Provide concrete guidance on the number of random draws that suffice to guarantee we learn a parameter region

Learning Framework

Abstract Definition of the Problem

- The parameters of interest are of the form $\lambda(\theta) \in \mathbb{R}^d$
  - $\theta \in \mathbb{R}^p$, parameters of a statistical model
  - λ function of interest; e.g., IRFs, subvector, identity
- $\lambda(\theta)$ belongs to a set, denoted $\lambda(S)$ - the parameter region of interest
  - $S \subseteq \mathbb{R}^p$
- Easy to check if $\theta \in S$: there is a labeling function $l : l(\theta) = 1\{\theta \in S\}$
- Difficult to describe $\lambda(S)$ analytically. All we know is $\lambda(S) \in \Lambda$

Sampling at Random: Supervised Learning Problem

- In order to evaluate $\lambda(S)$, use random sampling
  - Fix distribution P, generate i.i.d. $(\theta_1, \ldots, \theta_M)$
  - Use the draws and $I(\cdot)$ to report some set $\widehat{\lambda}_M$
- In supervised learning terminology (Mohri 2012)
  - $(\lambda(\theta_1), \ldots, \lambda(\theta_M))$ are *inputs* $\sim$ i.i.d. according to P
  - $(I(\theta_1), \ldots, I(\theta_M))$ are binary *labels* generated by $I(\theta) = \mathbf{1}\{\theta \in S\}$
  - $\widehat{\lambda}_M$ is a *learning algorithm*: a map from inputs and labels to Λ

Learning Criterion

- Define misclassification error

$$\mathcal{L}(\widehat{\lambda}_M; \lambda(S), P) \equiv P\left(\lambda(S) \triangle \widehat{\lambda}_M\right),$$

$\triangle$ is the symmetric set difference

DEFINITION 1: A parameter region $\lambda(S) \in \Lambda$ is said to be learnable if there exists an algorithm $\widehat{\lambda}_M$ and a number of draws $m(\epsilon, \delta)$ such that for any ϵ-δ:

$$P\left(\mathcal{L}(\widehat{\lambda}_M; \boldsymbol{\lambda}, P) < \epsilon\right) \geq 1 - \delta,$$

for all P on Θ and for any $\boldsymbol{\lambda} \in \Lambda$; provided $M \geq m(\epsilon, \delta)$.

Valiant, L.G. (1984): *A Theory of the Learnable.*

PAC-Learning.

Unpacking Learning

- Suppose there exists a journal referee that can compute misclassification events:

$$\lambda \in \lambda(S) \triangle \widehat{\lambda}_M$$

- The journal referee can P-compute how often misclassification happens:

$$\mathcal{L}(\widehat{\lambda}_M; \lambda(S), P) \equiv P\left(\lambda(S) \triangle \widehat{\lambda}_M\right),$$

- An approximation is ϵ-good when

$$\mathcal{L}(\widehat{\lambda}_M; \lambda(S), P) < \epsilon$$

Unpacking Learning

- The journal referee is worried
  1. That due to a insufficient number of draws, the quality of approximation may be poor too often
  2. About the distribution the econometrician uses
- To protect against this, the oracle requires

$$P\left(\mathcal{L}(\widehat{\lambda}_M; \lambda(S), P) < \epsilon\right) \geq 1 - \delta$$

Uniformly over all probability distributions and shapes of $\lambda(S)$.

What Can and Cannot be Learned

Learning $\iff$ VC Dimension of Λ is Finite

- THEOREM 1: $\lambda(S) \in \Lambda$ is learnable $\iff$ $\text{VC-dim}(\Lambda) < \infty$.

  PROOF: Blumer, Ehrenfeucht, Haussler, Warmuth (1989), *Learnability and the VC dimension*; Theorem 2.1

  VC dimension

- Assumptions to simplify econometric problems insufficient to learn (via random sampling)
  - Convex sets have infinite VC dimension $\implies$ cannot be learned
- If $\lambda(S)$ can't live in too complex of a class: what can we learn?

‘Tightest Bands’ for Parameter Regions

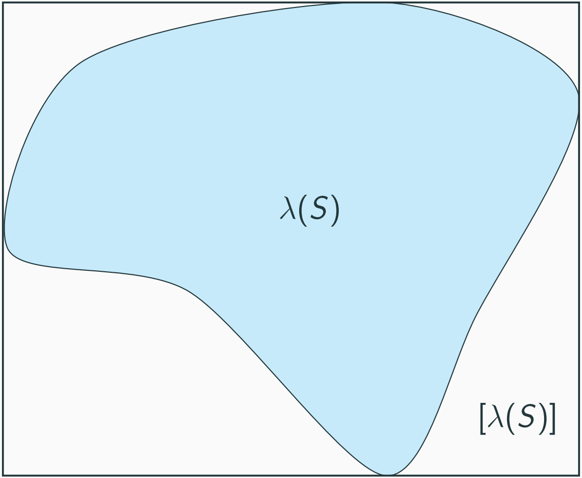
- Define the ‘tightest band’ containing $\lambda(S)$ as follows:

$$[\lambda(S)] \equiv \bigtimes_{j=1}^{d} \left[\inf_{\theta \in S} \lambda_j(\theta), \sup_{\theta \in S} \lambda_j(\theta)\right]$$

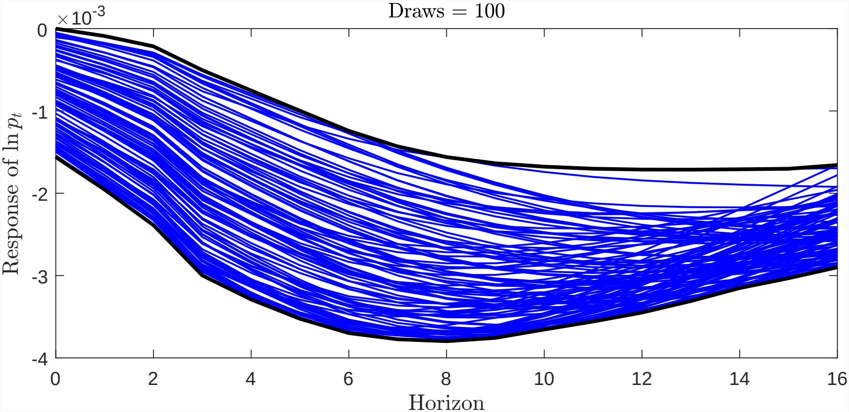
(λ_j is the j^{th} coordinate of $\lambda \in \mathbb{R}^d$)

- The tightest band is a d-dimensional axis-aligned hyperrectangle, which has VC dimension of $2d$

Tightest Bands that Contains $\lambda(S)$



Impulse Responses as Hyperrectangles



Algorithm for Learning Bands

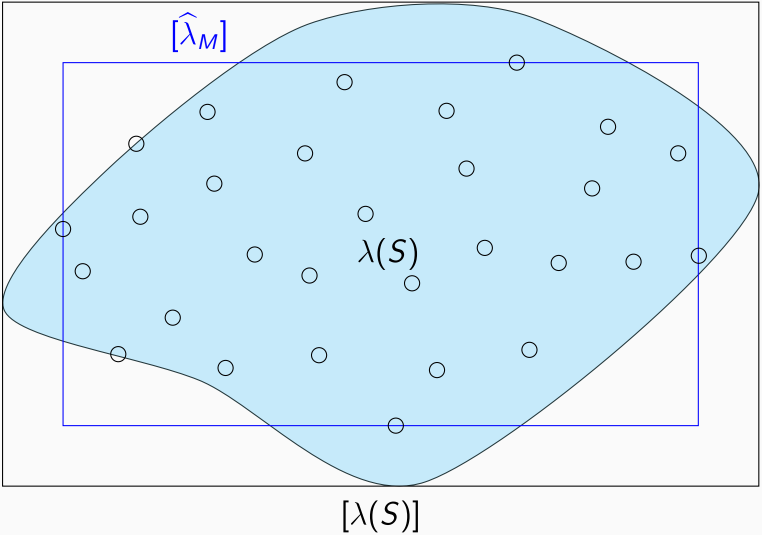
Definition 2: Given a sample $\boldsymbol{\theta}_M \equiv (\theta_1, \ldots, \theta_M)$ with labels $\boldsymbol{l}_M \equiv (l(\theta_1), \ldots, l(\theta_M))$, let $[\widehat{\lambda}_M]$ denote the algorithm that reports

$$[\widehat{\lambda}_M](\boldsymbol{\theta}_M, \boldsymbol{l}_M) := \bigtimes_{j=1}^{d} \left[\min_{m | l(\theta_m)=1} \lambda_j(\theta_m), \max_{m | l(\theta_m)=1} \lambda_j(\theta_m) \right]$$

where $\lambda_j(\theta)$ is the j-th element of $\lambda(\theta)$.

- Report min. and max. in each dimension (over positive labels).

Algorithm for Learning Bands



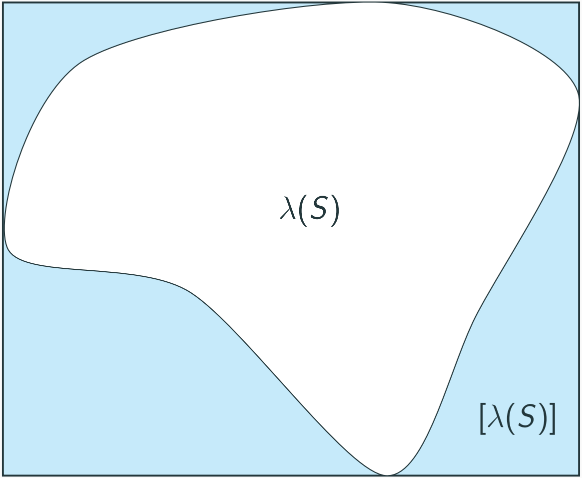
Can we Learn $[\lambda(S)]$ using $[\widehat{\lambda}_M]$?

- Is there a number of draws $m(\epsilon, \delta)$ such that

$$P\left(\mathcal{L}([\widehat{\lambda}_M]; [\boldsymbol{\lambda}], P) < \epsilon\right) \geq 1 - \delta$$

  for any P on Θ, for any $\boldsymbol{\lambda} \in \Lambda$, when $M > m(\epsilon, \delta)$?
- Theorem 2 in the paper answers this question in the negative ($\nexists$ algorithm that returns $\emptyset$ absent positive labels, and learns)
- Problem: allow for P's that put high mass on $[\lambda(S)] \setminus \lambda(S)$

Can we Learn $[\lambda(S)]$ using $[\widehat{\lambda}_M]$?



Learning $[\lambda(S)]$ from the Inside

- Define

$$\mathcal{P}(S) \equiv \{P \mid P \text{ is a distribution on } \Theta \text{ and } P(S) = 1\}$$

DEFINITION 3: The set $[\lambda(S)]$ is said to be learnable from the inside if there exists an algorithm $\widehat{\lambda}_M$ and a number of draws $m(\epsilon, \delta)$ such that

$$P\left(\mathcal{L}(\widehat{\lambda}_M; [\boldsymbol{\lambda}], P) < \epsilon\right) \geq 1 - \delta,$$

for any $P \in \mathcal{P}(S)$ and any $\boldsymbol{\lambda} \in \Lambda$, provided $M \geq m(\epsilon, \delta)$.

How Many Draws to Learn from the Inside?

Neccessary/Sufficient Draws to Learn $[\lambda(S)]$ from the Inside

THEOREM 3: The algorithm $[\widehat{\lambda}_M]$ learns $[\lambda(S)]$ from the inside. Moreover, the 'sample complexity' of $[\widehat{\lambda}_M]$—denoted $m^*(\epsilon, \delta)$—admits the following bounds:

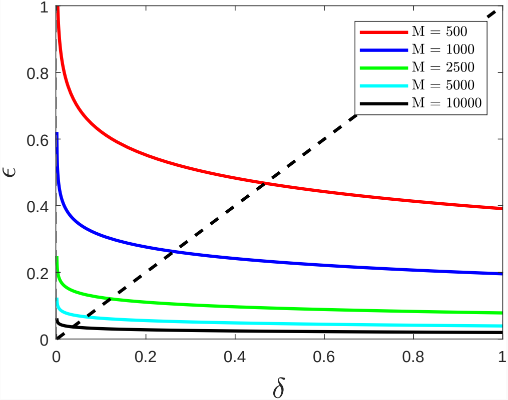
$$(1 - \epsilon/\epsilon) \ln(1/\delta) \leq m^*(\epsilon, \delta) \leq (2d/\epsilon) \ln(2d/\delta)$$

Proof

- For example $\epsilon = \delta = 0.01$, $d = 25$
  - Lower bound $= 456$
  - Upper bound $= 42,586$

Iso-Draw Curves

- If choosing ϵ-δ is a problem, commit to number of draws and report "Iso-Draw curves"
  - All the ϵ-δ that support upper bound on sample complexity



Example (Revisited)

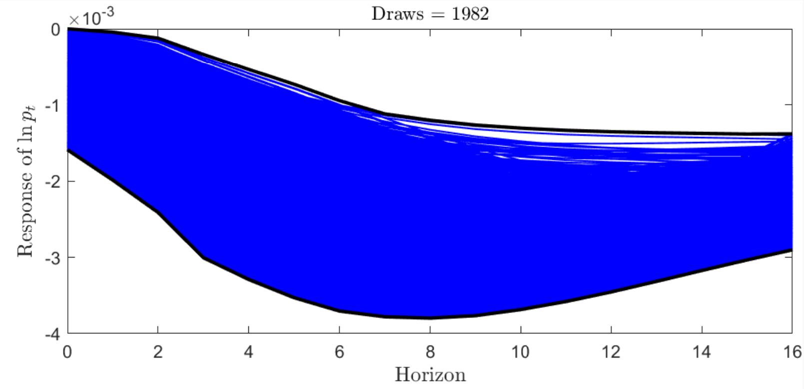
$$d = 17 \text{ (impact + 16 quarters)}$$

$$\epsilon = \delta = 0.1$$

Numbers of draws from inside sufficient for learning:

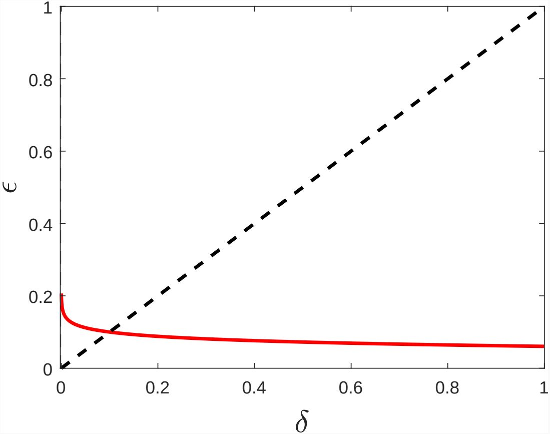
$$\frac{2d}{\epsilon} \log\left(\frac{2d}{\delta}\right) = 1,982$$

Example (Revisited)



Misclassifcation error of less than 10% with probability at least 90%

Example (Revisited)



Iso-draw curve: The values of ϵ-δ that can be supported with $1,982$ draws.

Conclusion

Conclusion

- Random sampling approximations are supervised learning problems
- Misclassification error and learning are natural criterion to judge the accuracy of these approximations
- Learning a parameter region is possible iff the class is not too complex
- Defined learning from the inside and showed that to learn the tightest bands from the inside $(2d/\epsilon)\ln(2d/\delta)$ draws suffice

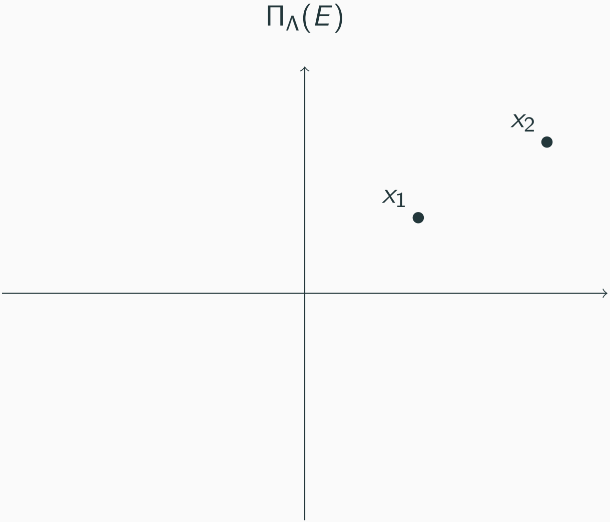
Thank you for listening!

VC Dimension

- Let E denote a finite subset of $\mathbb{R}^d$ (the space in which $\boldsymbol{\Lambda}$ lives)
- Define $\Pi_{\boldsymbol{\Lambda}}(E) \equiv \{E \cap \Lambda \mid \Lambda \in \boldsymbol{\Lambda}\}$
- Say that E is *shattered* by the class $\boldsymbol{\Lambda}$ whenever $\Pi_{\boldsymbol{\Lambda}}(E) = 2^E$
- Define the VC dimension of $\boldsymbol{\Lambda}$ as the cardinality of the largest finite set of points E shattered by the class $\boldsymbol{\Lambda}$
- Vapnik, V. (1998): *Statistical Learning Theory*
- Axis-aligned rectangles have a VC dimension of 4

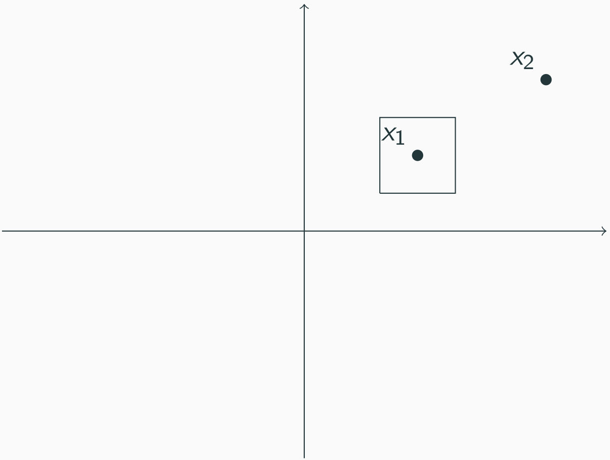
Back

$E = \{x_1, x_2\}$; **Λ : All axis-aligned rectangles**



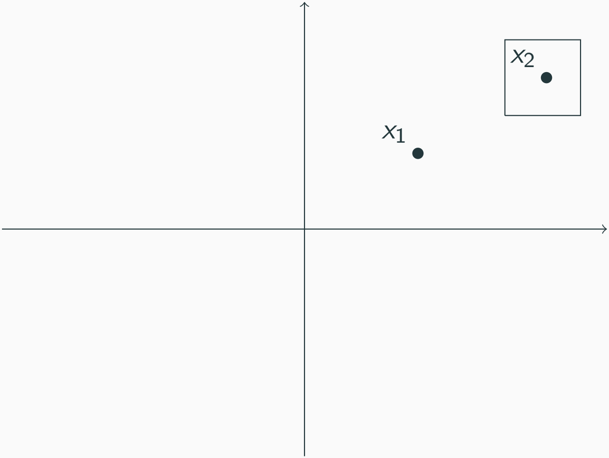
$E = \{x_1, x_2\}$; Λ : All axis-aligned rectangles

$$E \cap \Lambda = \{x_1\}$$



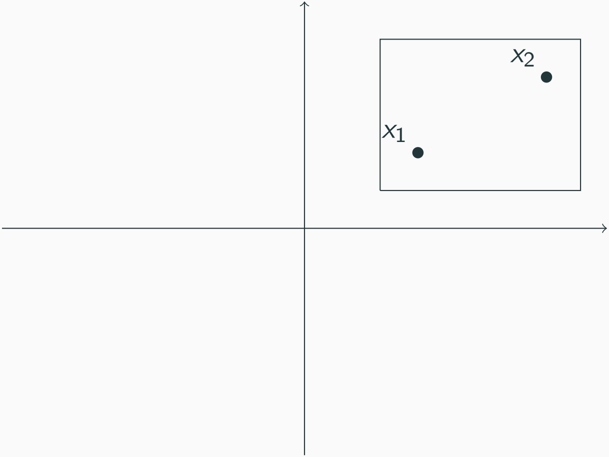
$E = \{x_1, x_2\}$; Λ : All axis-aligned rectangles

$$E \cap \Lambda = \{x_2\}$$



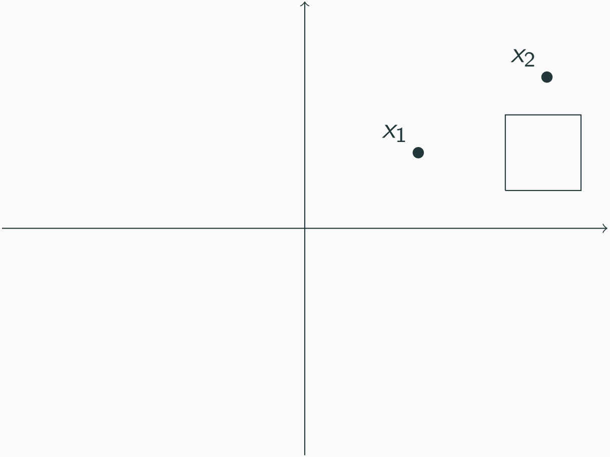
$E = \{x_1, x_2\}$; **Λ : All axis-aligned rectangles**

$$E \cap \Lambda = \{x_1, x_2\}$$



$E = \{x_1, x_2\}$; Λ : All axis-aligned rectangles

$$E \cap \Lambda = \{\emptyset\}$$



$$\Pi_\Lambda(E) = \{\{x_1\}, \{x_2\}, \{x_1, x_2\}, \{\emptyset\}\}$$

VC-dim $\boldsymbol{\Lambda} \geq 2$

(In fact VC-dim $\boldsymbol{\Lambda} = 4$)

Back

Proof of Theorem 3

- The true concept is $[\lambda(S)]$ and the estimator is $[\widehat{\lambda}_M]$
- Note that for any draws we have that $[\widehat{\lambda}_M] \subseteq [\lambda(S)]$. Consequently:

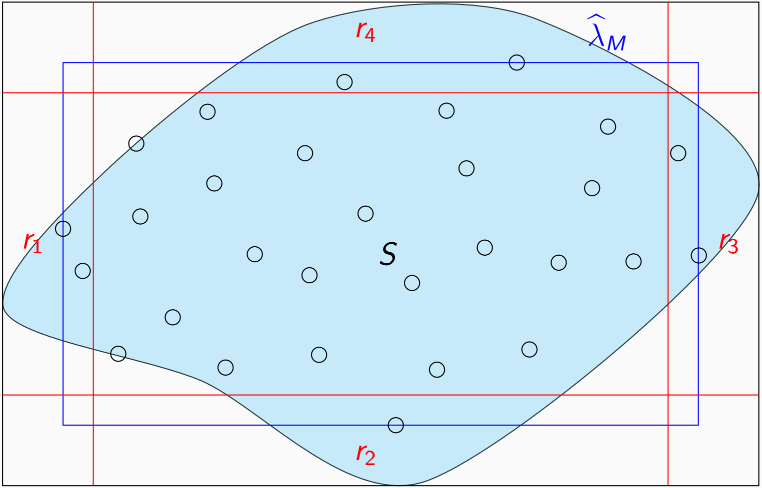
$$\mathcal{L}([\widehat{\lambda}_M]; [\lambda(S)], P) = P(\theta \in [\lambda(S)] \setminus [\widehat{\lambda}_M])$$

- Since P samples from inside S, then $P(\lambda(\theta) \in [\lambda(S)]) = 1$.
- Construct $2d$ 'special hyperrectangles' $(r_1, r_2, \cdots, r_{2d})$, parallel to each side of $[\lambda(S)]$, each with probability $\geq \epsilon/2d$ (but interior has probability $\leq \epsilon/2d$)

Proof of Theorem 3

- Consider the event that $[\widehat{\lambda}_M]$ intersects each of these $2d$ special rectangles
- The misclassification error is less than the probability of the union of the interior of these rectangles

$$\mathcal{L}([\widehat{\lambda}_M]; [\lambda(S)], P) \leq \sum_{j=1}^{2d} \epsilon/2d = \epsilon$$



$[\lambda(S)]$

Proof of Theorem 3

$$
\begin{aligned}
P(\mathcal{L}([\widehat{\lambda}_M]; [\lambda(S)], P) > \epsilon) &\leq P([\widehat{\lambda}_M] \cap r_j = \emptyset, \text{ for some } j) \\
&\leq \sum_{j=1}^{2d} P([\widehat{\lambda}_M] \cap r_j = \emptyset) \\
&= \sum_{j=1}^{2d} P(\lambda(\theta_m) \notin r_j \text{ or } \theta_m \notin S)^M \\
&\leq \sum_{j=1}^{2d} P(\lambda(\theta_m) \notin r_j)^M \\
&\leq 2d(1 - \epsilon/2d)^M
\end{aligned}
$$

Proof of Theorem 3

Learning is guaranteed whenever

$$2d(1-\epsilon/2d)^M \le \delta$$

$$\iff$$

$$M \ge -\ln(2d/\delta)/\ln(1-\epsilon/2d)$$

Roughly

$$M \ge \frac{2d}{\epsilon}\ln\left(\frac{2d}{\delta}\right)$$

Back