Latin Modern Math

The Bridge Between Theory and Applications

An Examination of Bias in Nonlinear Classification

Travis McVoy

April 19, 2025

Binary Classification

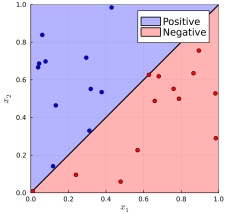


Figure: Linear Classifier

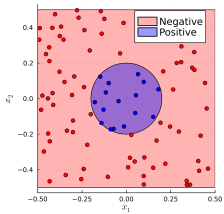


Figure: Loop Classifier

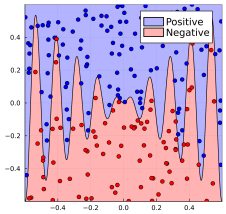
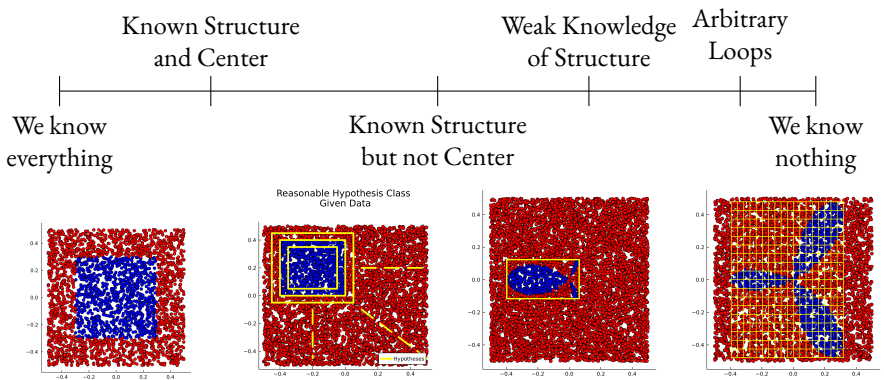


Figure: Curve Classifier

Motivating Questions and Ideas

- No Free Lunch
  - Provably impossible to find universal learner
  - We need bias to learn
- How much bias is enough?
  - For linear data, knowledge of linear structure is enough
  - What about nonlinear data?



Learning Theory Definitions

- Learning space is denoted by $\mathcal{X}$
  - Examples:
    - Cartesian plane $\mathbb{R}^2$
    - Three Space $\mathbb{R}^3$,
    - Subspaces $[-0.5, 0.5]^2$
- Set of labels is denoted by $\mathcal{Y}$
  - For binary classification $\mathcal{Y} = \{0, 1\}$
- Training data is denoted by a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Labeling function: $f : \mathcal{X} \to \mathcal{Y}$

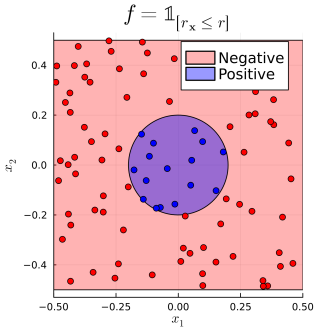


Figure: Visual Summary of Definitions

Empirical Risk Minimization (ERM)

True Risk:

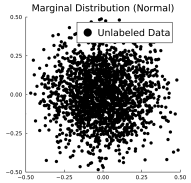
- Probability of Misclassifying an Instance
- "Calculated" with Marginal Distribution $\mathcal{D}_{\mathcal{X}}$ and true labeling function f

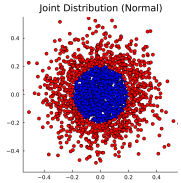
$$\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}[f(\mathbf{x}) \neq h(\mathbf{x})]$$

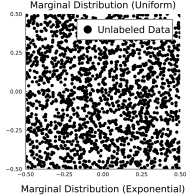
Empirical Risk:

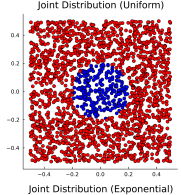
- Ratio of Mislabeled Points to All Points
- Calculated with Sample

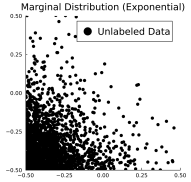
$$\frac{1}{|S|}\sum_{i=1}^{|S|}\mathbb{1}_{[f(\mathbf{x}_i)\neq h(\mathbf{x}_i)]}$$

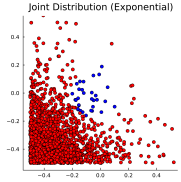






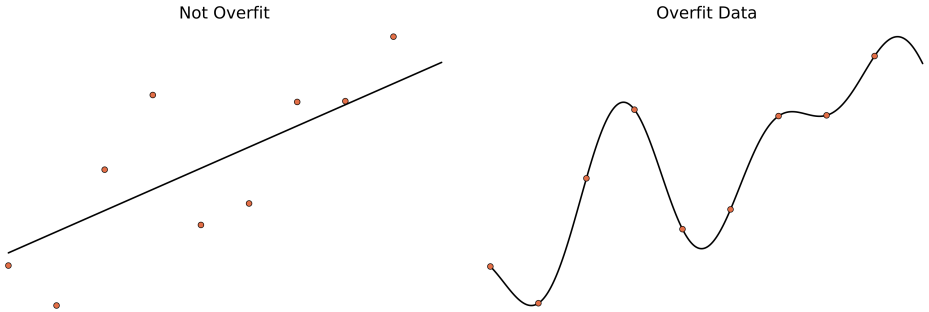






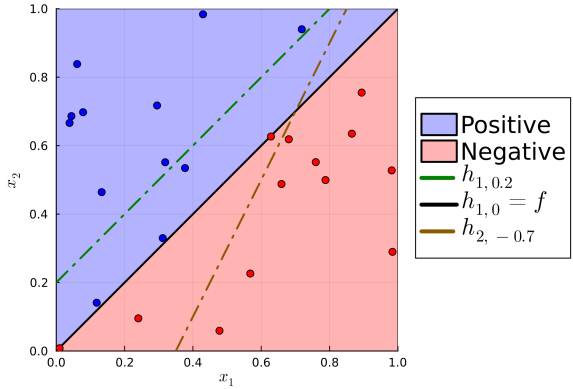
Overfitting

ERM can lead to overfitting (poor generalization)



Hypothesis Classes

- Learning model chooses from set of functions $\mathcal{H}$
- Each function is of from $h : \mathcal{X} \to \mathcal{Y}$
- We call h a hypothesis
- We call $\mathcal{H}$ a hypothesis class
  - Classes enable us to induce bias into models
  - Example: Linear Classifiers
    - $\mathcal{H} = \{h_{m,b} : m, b \in \mathbb{R}\}$
    - $h_{m,b} = \mathbb{1}_{[x_2 \geq m x_1 + b]}$



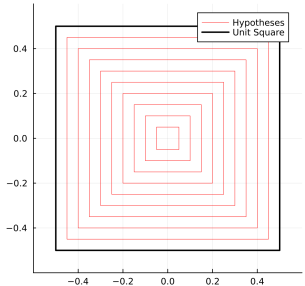
Probably Approximately Correct (PAC) Learnability

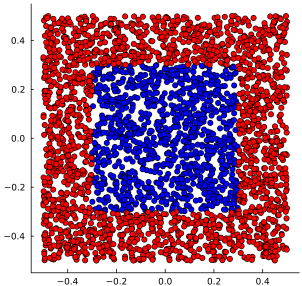
- Realizability Assumption
  - There exists $h^* \in \mathcal{H}$ such that the true risk of h^* is zero
  - If assumption holds, ERM is a probably (with confidence $1 - \delta$) approximately (up to an error ϵ) correct learner
  - Example:
    - Set $\epsilon = 0.1$ and $\delta = 0.2$
    - Consider 1000 trials
    - If S sufficiently large, ≥ 800 trials should result in $\geq 90\%$ accuracy
- Important: realizable learners are arbitrarily strong
- What if realizability fails?
  - Agnostic PAC learning
    - An agnostic $\mathcal{H}$ is only so strong
    - That is, there exists an $h \in \mathcal{H}$ for which the risk is minimal but nonzero

Sample Complexity

- How large does S have to be for good results? Depends
  - If $\mathcal{H}$ finite and realizable, sample complexity is $n_{\mathcal{H}}(\epsilon, \delta) = \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$
  - If $\mathcal{H}$ is finite and agnostic, sample complexity is $n_{\mathcal{H}}(\epsilon, \delta) = \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$
  - If $\mathcal{H}$ is infinite, ERM only works if $VC(\mathcal{H}) < \infty$
- The above is not as important as the following:
  - In practice, most models will be agnostic
  - Agnostic classes vary in strength, so data is not really the issue

Concentric Squares at the Origin



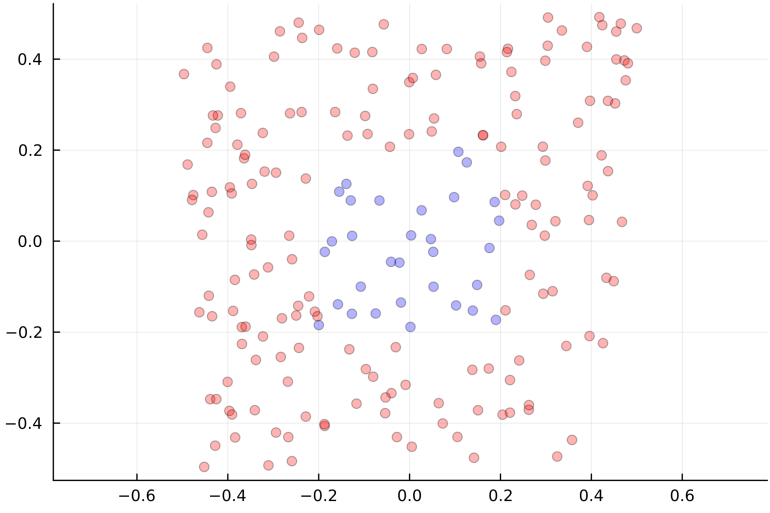


Known Structure and Center

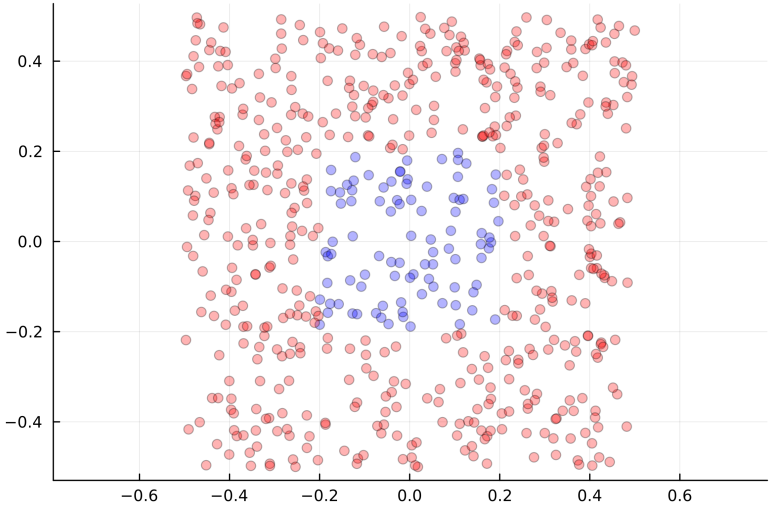
We know everything

We know nothing

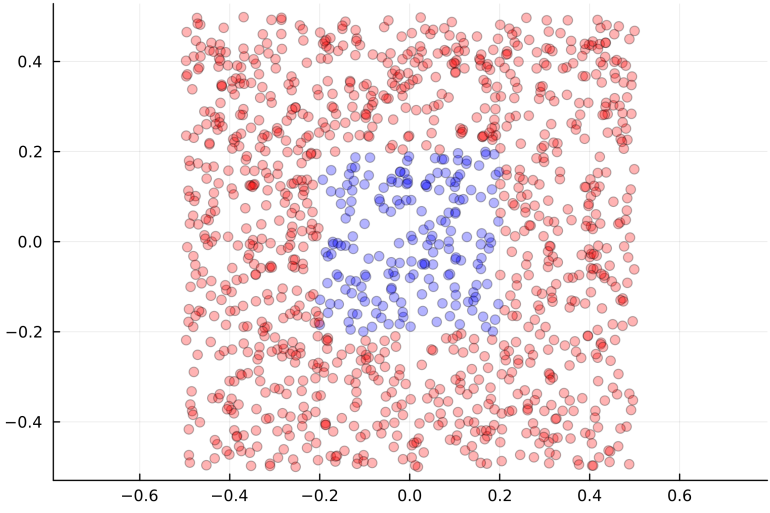
Monte Carlo Simulations



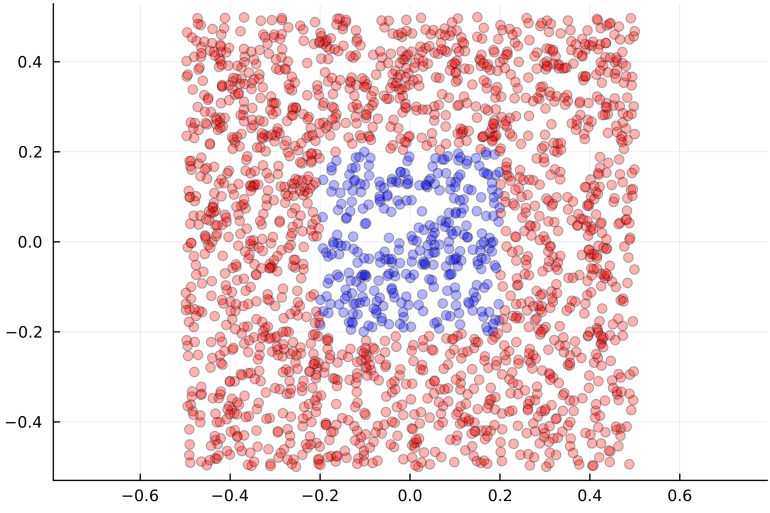
Monte Carlo Simulations



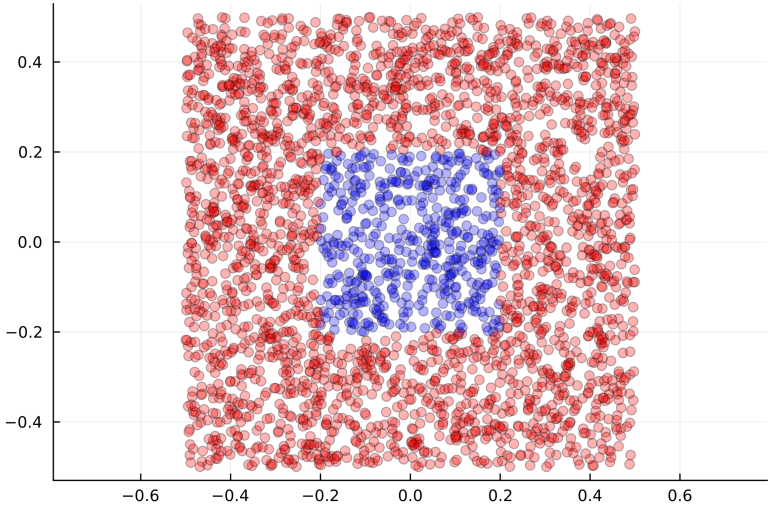
Monte Carlo Simulations



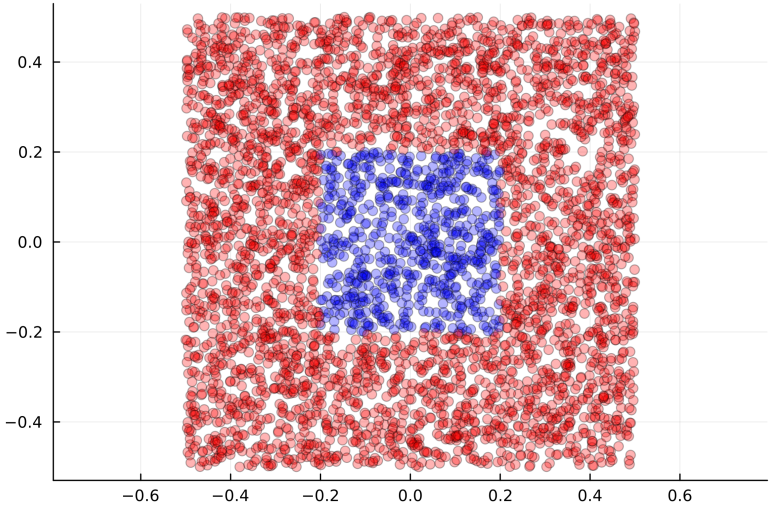
Monte Carlo Simulations



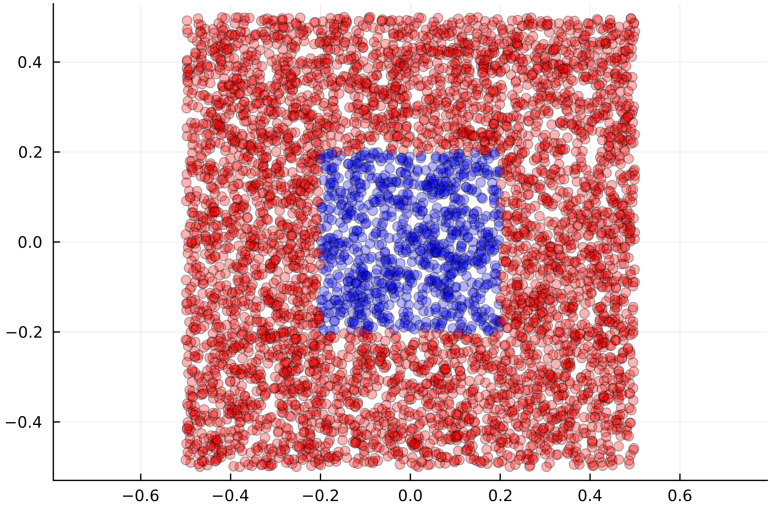
Monte Carlo Simulations



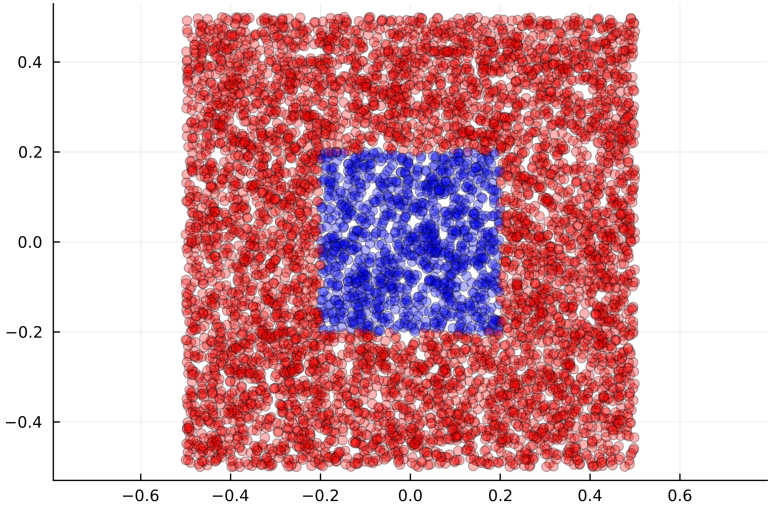
Monte Carlo Simulations



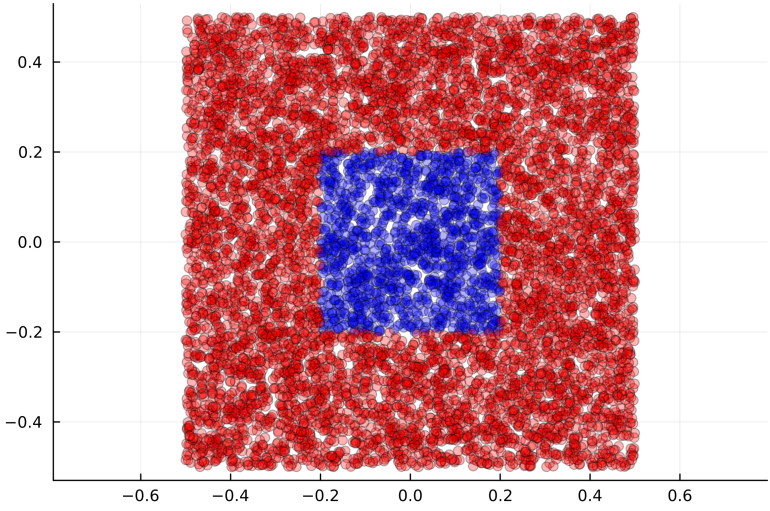
Monte Carlo Simulations



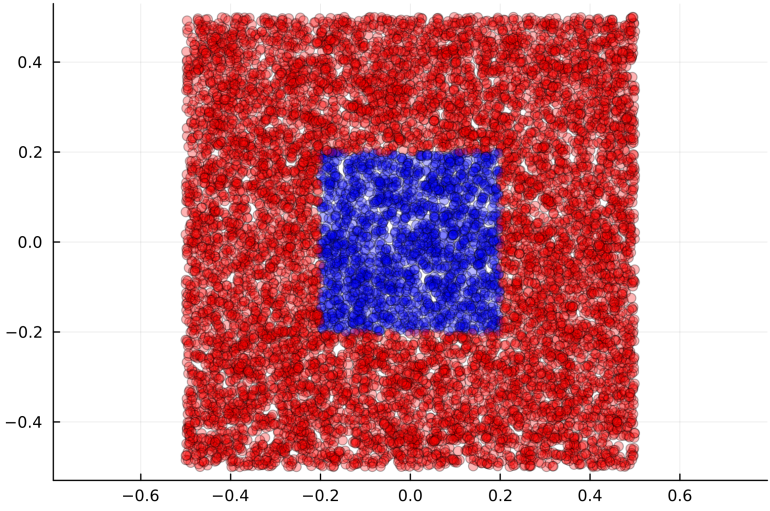
Monte Carlo Simulations



Monte Carlo Simulations



Monte Carlo Simulations

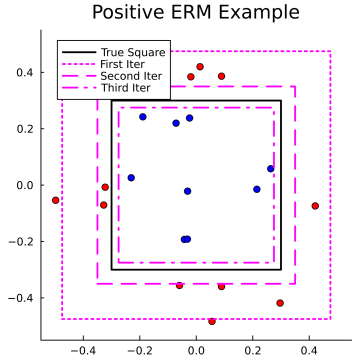


Training Concentric Squares at the Origin

Three Algos: Positive, Negative, Midpoint

Algorithm 1 Positive ERM

1: Initialize square Q to be largest square possible
2: **for** each instance in S **do**
3: **if** Q mislabels **x** **then**
4: Decrease Q such that mislabel is corrected
5: **end if**
6: **end for**

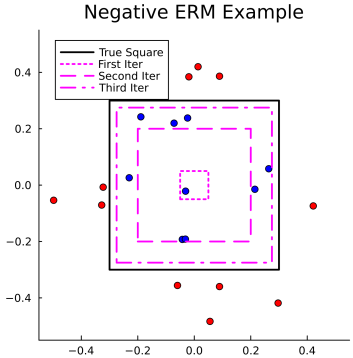


Training Concentric Squares at the Origin

Three Algos: Positive, Negative, Midpoint

Algorithm 2 Negative ERM

1: Initialize square Q to be smallest square possible
2: **for** each instance in S **do**
3: **if** Q mislabels **x** **then**
4: Increase Q such that mislabel is corrected
5: **end if**
6: **end for**

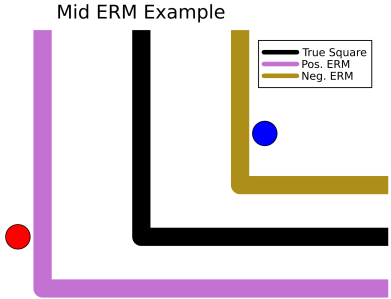


Training Concentric Squares at the Origin

Three Algos: Positive, Negative, Midpoint

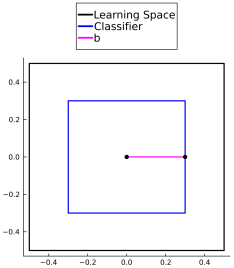
Algorithm 3 Midpoint ERM

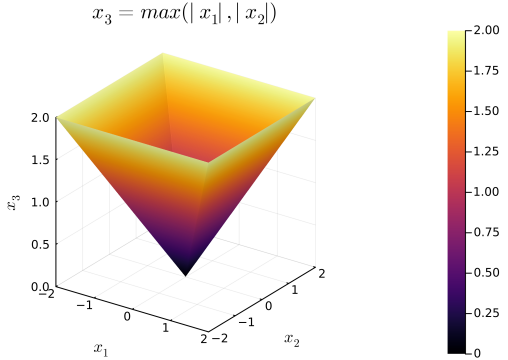
1: Let Q_P be square from positive ERM
2: Let Q_N be square from negative ERM
3: **return** the square in the middle of Q_N and Q_P



Mid ERM Achieves Kernel-SVM

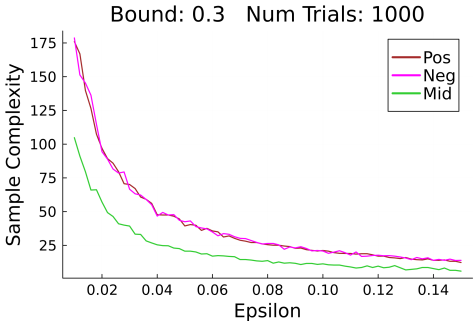
- Support Vector Machines (SVM) maximize margin between halfspaces
  - Reduces sample complexity
- Hard SVMs require data to be linearly separable, but our data is not, nor did we transform it

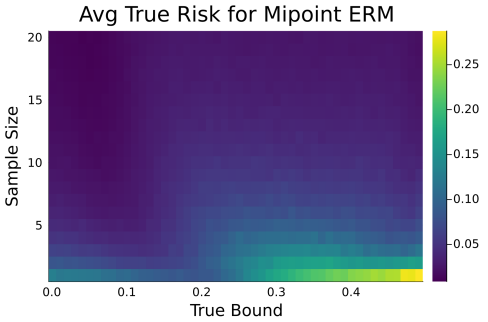




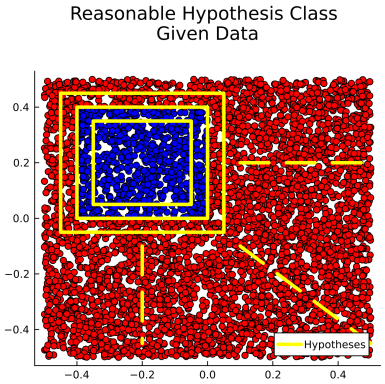
Experiment Results

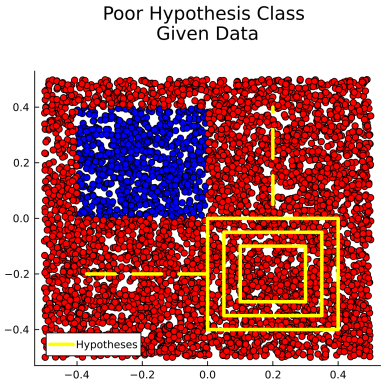
Experiments with $\delta = 0.1$





Concentric Squares with Unkown Center





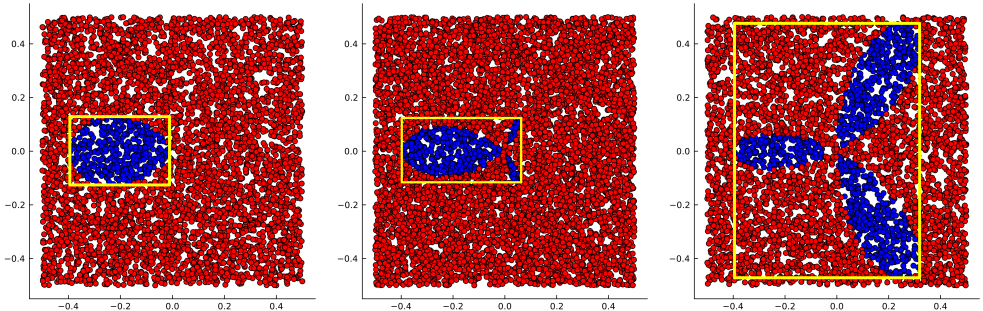
Known Structure
and Center

We know
everything

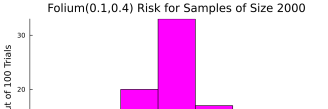
Known Structure
but not Center

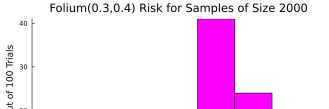
We know
nothing

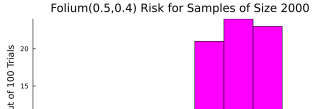
Rectangular Approximations

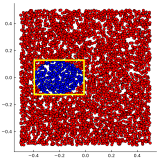


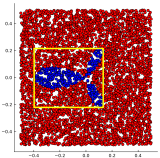
Rectangles Are Only So Strong

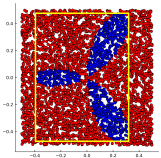






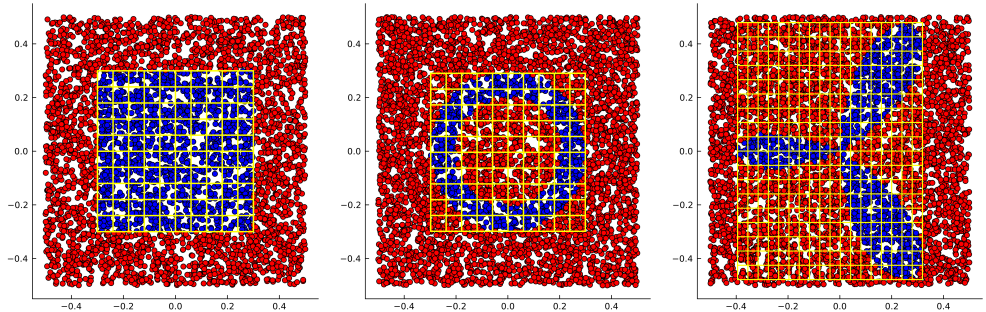






- Main point: Rectangles are agnostic, and their strength is conditioned on loop structure, not sample size

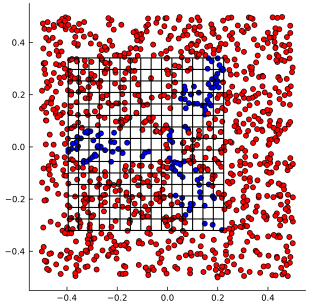
Grid Classifiers



Algorithm 4 Scan Grid

1: Initialize a hash table H to be empty.
2: Let S be our sample, B our border, $\mathbf{G}$ our untrained grid matrix.
3: **for** each instance $\mathbf{x} \in S$ **do**
4: **if** $\mathbf{x} \in B$ **then**
5: Locate the indices i, j such that $\mathbf{x} \in \mathbf{G}_{i,j}$
6: **if** $(i, j) \notin H$ **then**
7: add (i, j) to H
8: **end if**
9: **end if**
10: **end for**
11: **return** $\mathbf{G}$, H

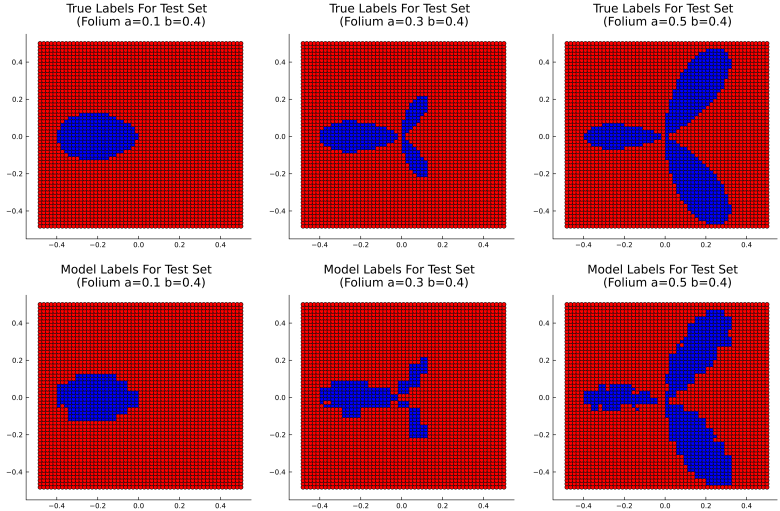
- Alg. 4 is both how we train fixed size and determine grid size



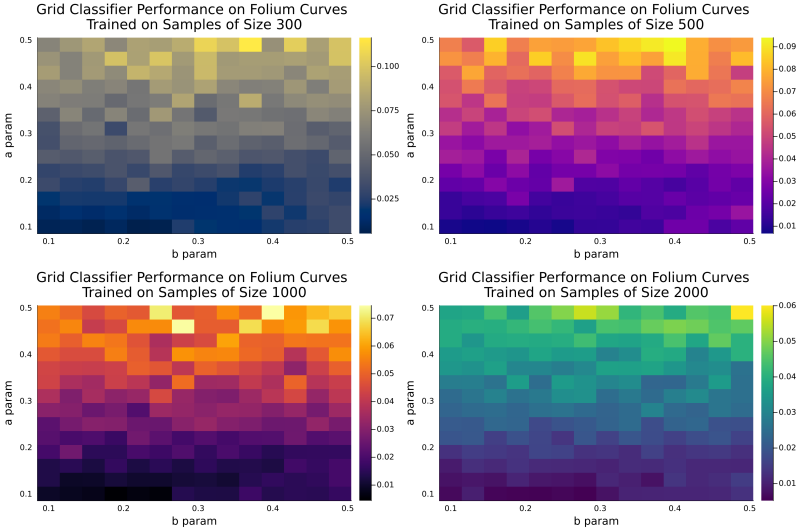
- If scan returns grid with empty cell, decrease size. Otherwise increase.
- While scan occurs, we update grid matrix

How Strong Are Grid Classifiers?

Results for Large (several thousand instances) Data Sets

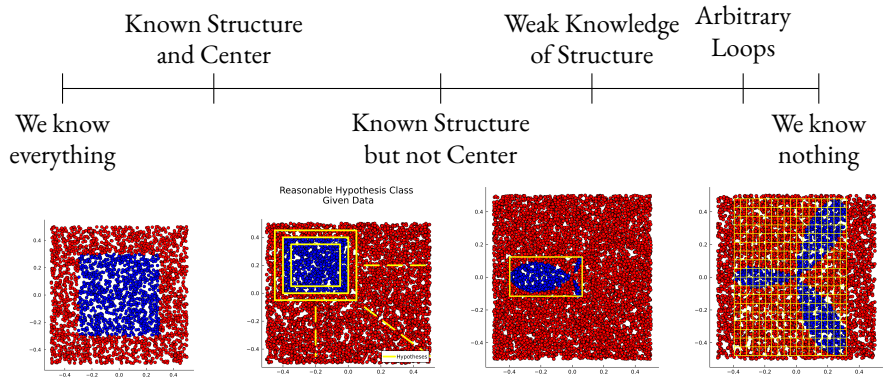


Can Grid Classifiers Learn With Small Data Sets?



Conclusion

- **Main Question:** Knowledge of Linear Structure is enough for Linear Classification; are the nonlinear analogues?
- **Answer:** Seems like it!



References

1. Valiant, L. G. A theory of the learnable. *Commun. ACM* **27**, 1134–1142. doi:10.1145/1968.1972 (1984).
2. Wolpert, D. & Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67–82. doi:10.1109/4235.585893 (1997).
3. Vapnik, V. An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10**, 988–999. doi:10.1109/72.788640 (1999).
4. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. (Cambridge University Press, USA, 2014).
5. Sharan, V. *University of Southern California Theory of Machine Learning*. https://vatsalsharan.github.io/fall21.html. Accessed: 2025-04-15. 2021.
6. Schapire, R. E. *Princeton's COS 511: Theoretical Machine Learning*. https://www.cs.princeton.edu/courses/archive/spr08/cos511/schedule.html. Accessed: 2025-04-15. 2008.
7. Royden, H. L. & Fitzpatrick, P. *Real analysis / H.L. Royden, Stanford University, P.M. Fitzpatrick, University of Maryland, College Park.*. Fourth edition [2018 reissue]. eng (Pearson, New York, NY, 2018 - 2010).
8. Jkasd. *Knot table*. Accessed: 2025-04-15. Licensed under CC BY-SA 3.0 Unported. 2008.
9. Barnett, J. H., Adams, P., *et al. Fifty Famous Curves*. Accessed: 2025-04-16. 1999.
10. Tttrung. *Klein bottle*. Image licensed under the GNU Free Documentation License. Accessed: 2025-04-16. 2005.
11. Munkres, J. R. *Topology*. 2nd ed. (Prentice Hall, Inc., 2000).
12. Rice, J. A. *Mathematical Statistics and Data Analysis*.. Third (Belmont, CA: Duxbury Press., 2006).
13. Asilis, J., Devic, S., *et al. Open Problem: Can Local Regularization Learn All Multiclass Problems?*. in *Proceedings of Thirty Seventh Conference on Learning Theory* (eds Agrawal, S. & Roth, A.) **247** (PMLR, 2024), 5301–5305.
14. Brukhim, N., Carmon, D., *et al. A Characterization of Multiclass Learnability*. in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Computer Society, Los Alamitos, CA, USA, 2022), 943–955. doi:10.1109/FOCS54457.2022.00093.
15. Dughmi, S., Kalayci, Y. & York, G. Is Transductive Learning Equivalent to PAC Learning? *arXiv preprint arXiv:2405.05190v2*. Accessed: 2025-04-17 (2024).