

How Boosting the Margin Can Also Boost Classifier Complexity

ICML '06

Lev Reyzin

Yale University

Robert Schapire

Princeton University



The Learning Task

- Given m training examples and their labels
- Predict the label of a new example



The Idea of Boosting

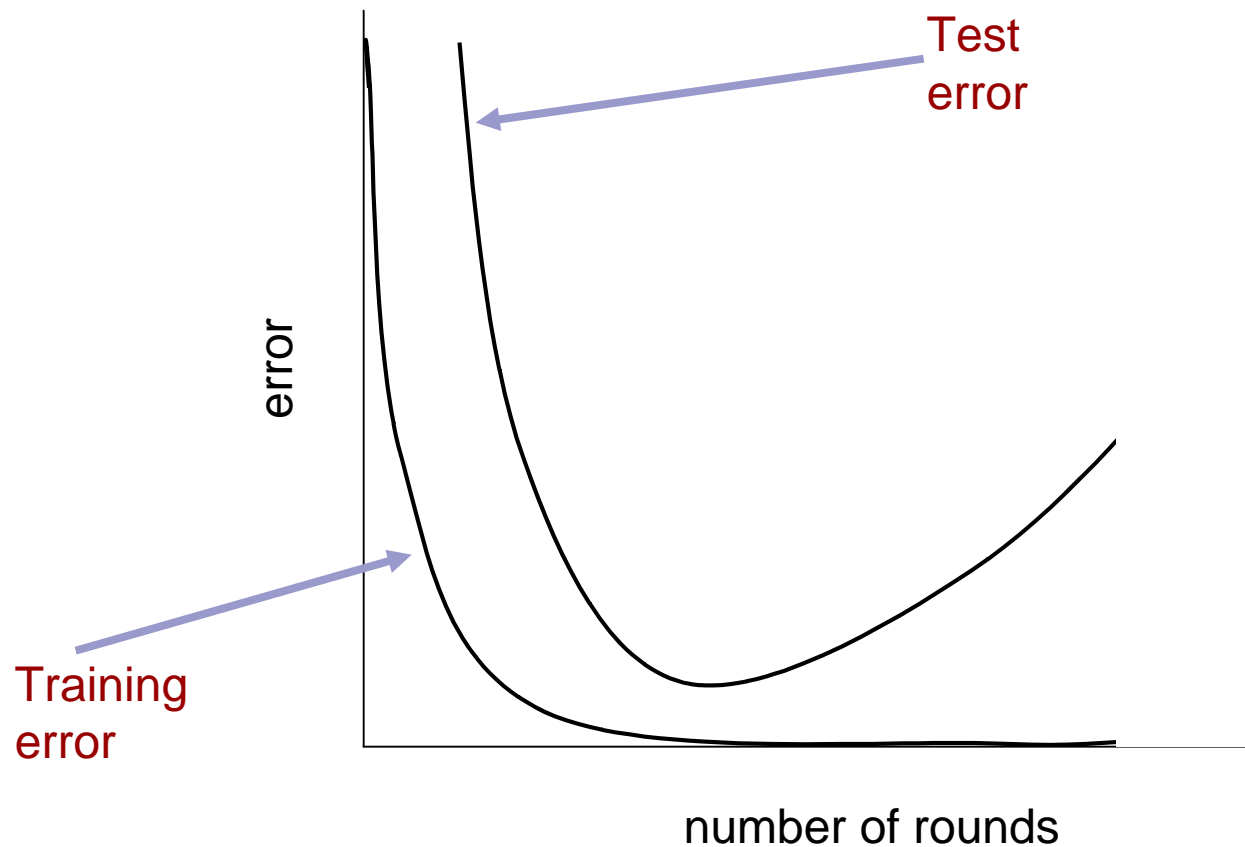
- Combine many “moderately inaccurate” **base classifiers** into a combined predictor
- Generate a new base classifier in each round
- Constantly focus on the **hardest** examples
- The final predictor is the **weighted vote** of the base classifiers
- **AdaBoost** sets voting weights of each new base classifier to reduce an upper bound on the training error.



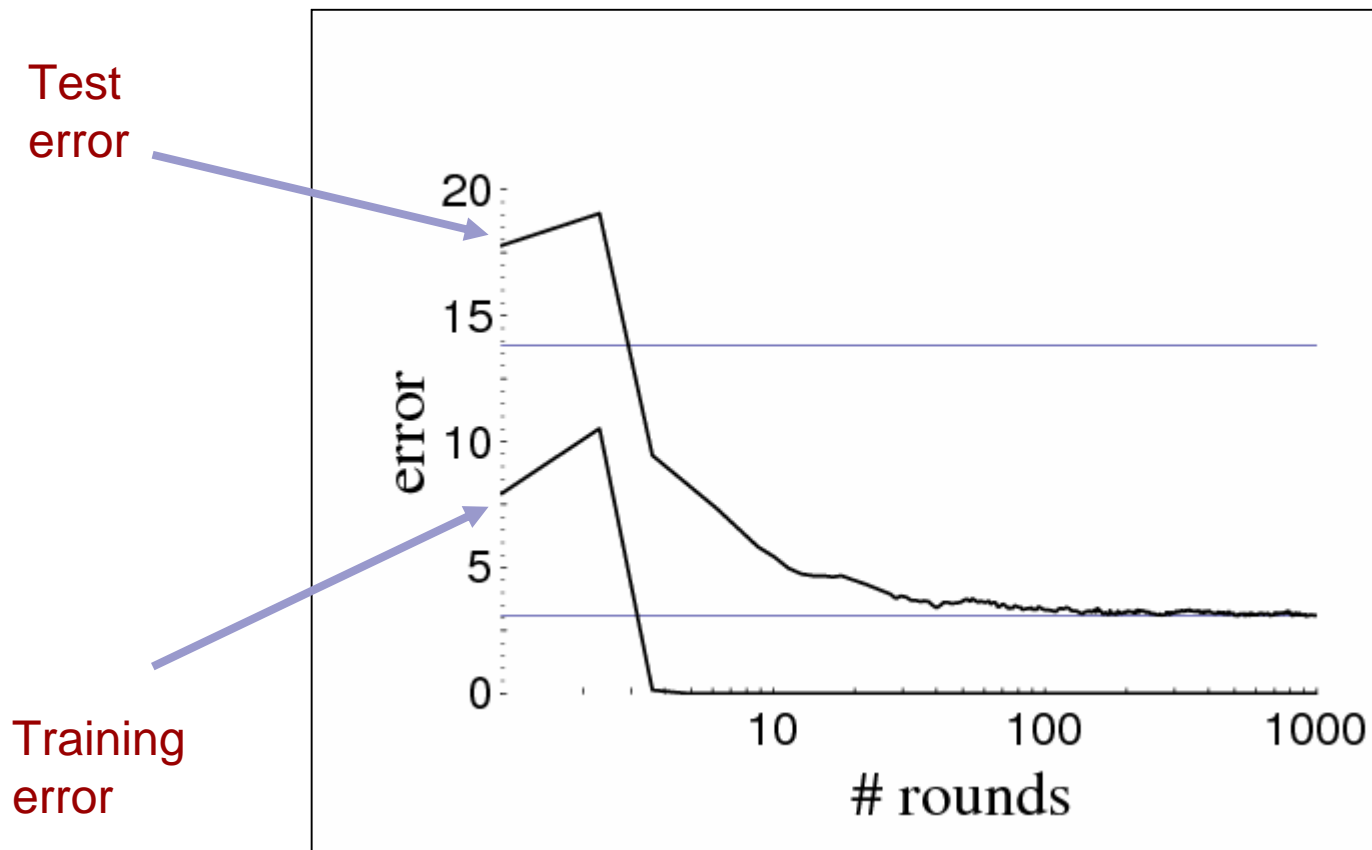
An Early Bound on Generalization Error

- An early generalization error bound depended on the number of boosting rounds. [Freund & Schapire '97]
- This meant that AdaBoost should “overfit” the training data.

We Would Expect Overfitting



However...



[Drucker & Cortes; Breiman; Quinlan, ...]



The Margin

- The margin of a classifier on an example:
 - margin = (weighted fraction of base classifiers voting for correct label) – (weighted fraction voting for incorrect label)
 - magnitude represents the **confidence** of the vote
 - positive if the vote gives the correct classification. Otherwise it's negative.

- **Margins** are measured over training examples

A Margin Bound

- A later bound relied on the margins the classifier achieved on the training examples and not on the number of rounds of boosting. [Schapire et. al. '98]

the generalization error is at most:

$$\hat{\Pr} \left[\text{margin}_f(x, y) \leq \theta \right] + \tilde{O} \left(\sqrt{\frac{d}{m\theta^2}} \right)$$

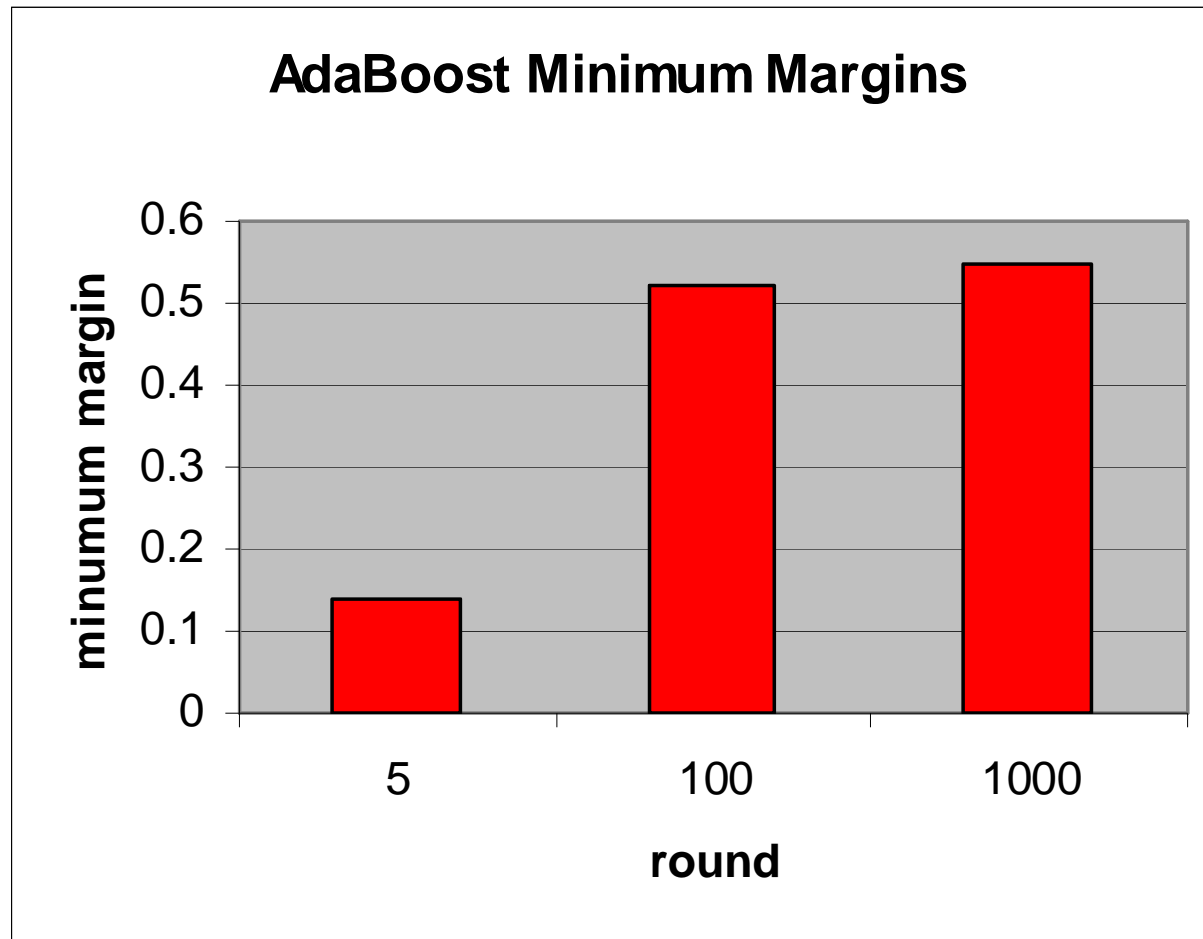
Fraction training examples with margin below theta

number of training examples

for any value of theta

the VC dimension of the base classifier

AdaBoost's Minimum Margins





The Margins Explanation

- AdaBoost pushes the cumulative margins distribution towards higher margins.
- All things being equal, **higher margins mean lower generalization error.**



arc-gv [Breiman '98]

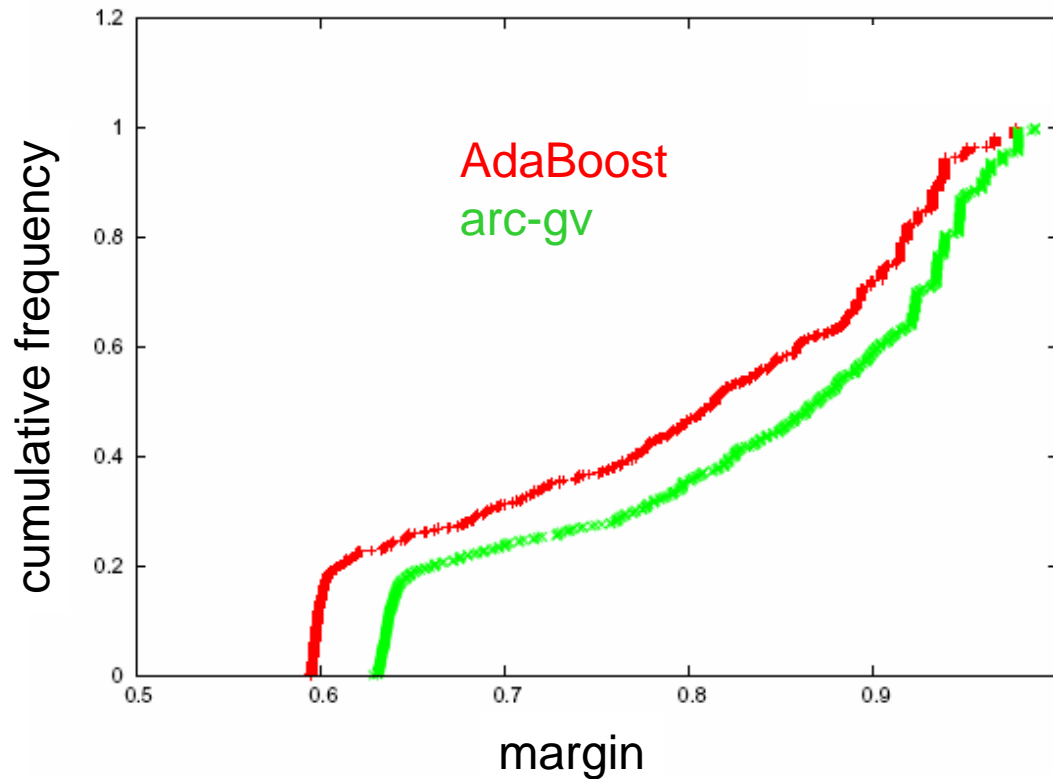
- motivated by the margins explanation
 - arc-gv's **minimum margin** provably converges to the **optimal**
 - one line difference from AdaBoost
- Breiman's reasoning: higher minimum margin would imply lower test error



The Experiments

- Data: Breast cancer, ionosphere, and splice
 - From UCI
 - Same natural datasets as Breiman used
- Data: ocr 17, ocr 49
 - Random subsets from NIST
 - Scaled to 14x14 points
- Binary classification
- Use 16-leaf CART trees as base classifiers

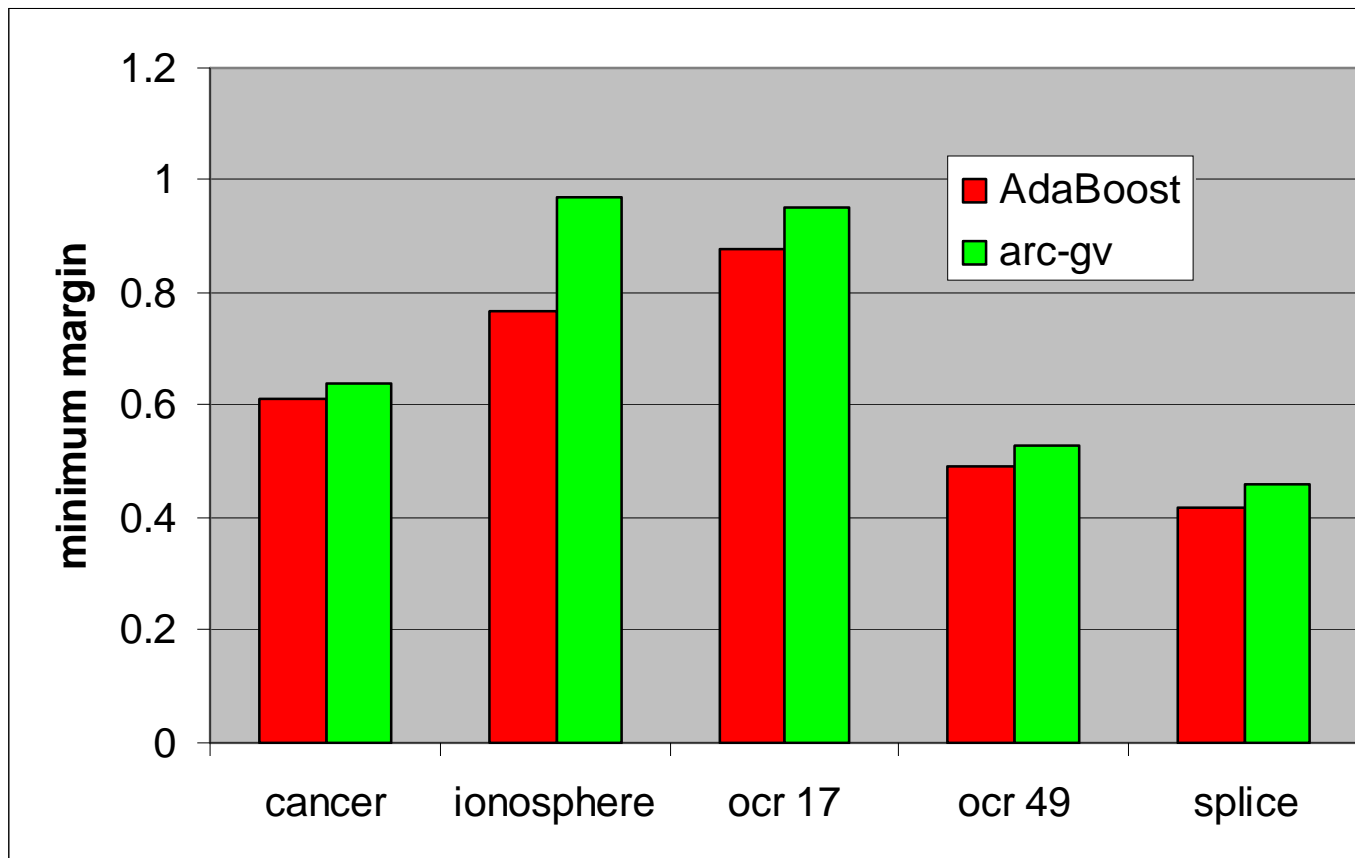
Data: the Margins



Cumulative margins: 500 rounds of boosting on the “breast cancer” dataset using pruned CART trees as weak learners.

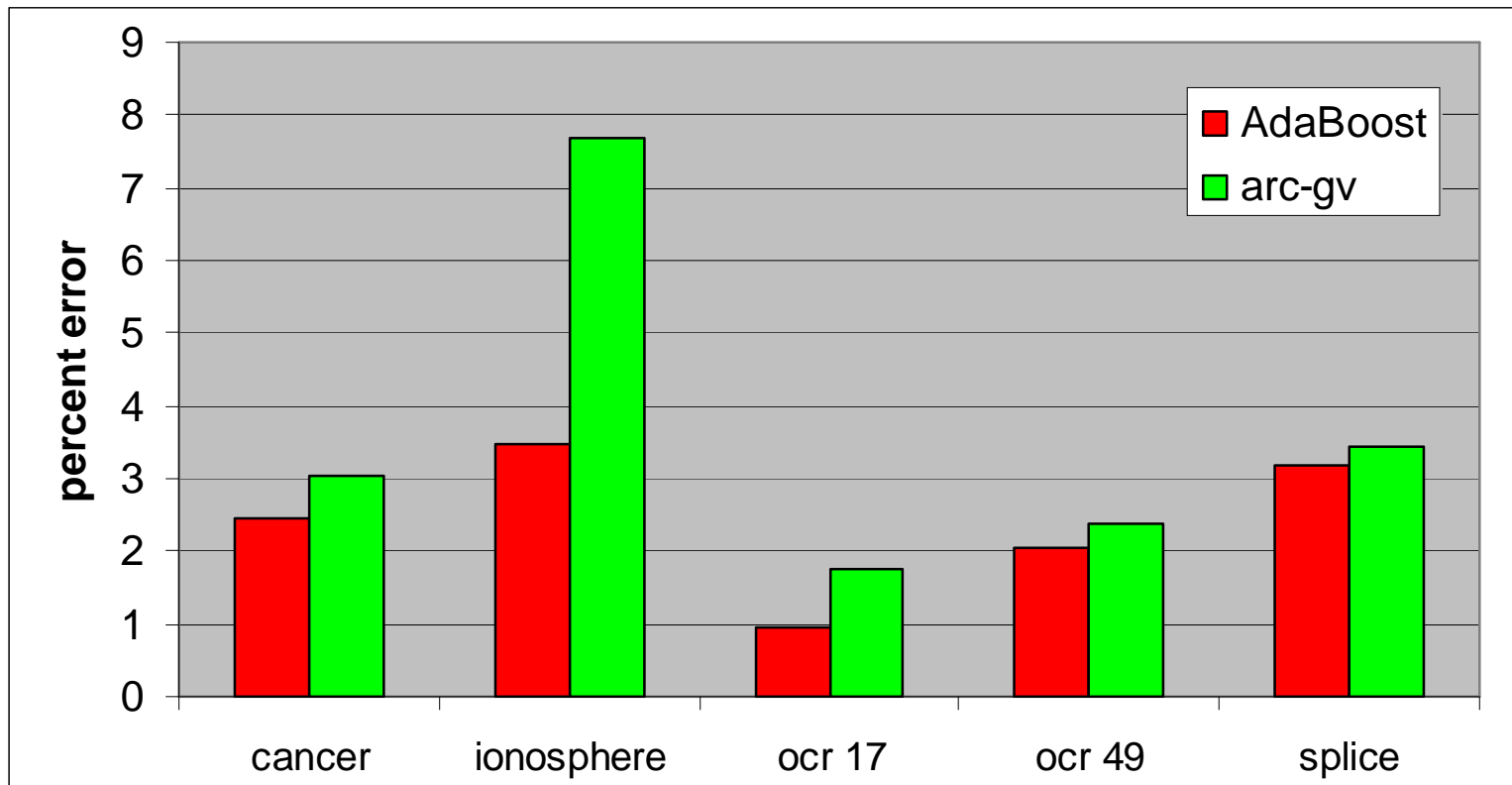
The Minimum Margins

Minimum margins of AdaBoost and arc-gv with pruned CART trees as base classifiers



Data: the Errors

Test errors of AdaBoost and arc-gv with pruned CART trees as base classifiers





Doubting the Margins Explanation

- arc-gv has **uniformly higher** margins than AdaBoost with pruned CART trees.
- the margins explanation predicts that arc-gv should perform better, but instead arc-gv performs worse.
- Breiman's experiment put the margins theory into serious doubt



Reconciling with Margins Theory?

$$\hat{\Pr} [\text{margin}_f(x, y) \leq \theta] + \tilde{O} \left(\sqrt{\frac{d}{m\theta^2}} \right)$$

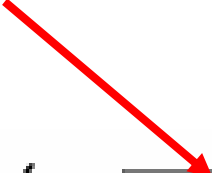
- Margin bound depends on the entire distribution – not just minimum margin.
 - But arc-gv's margins were uniformly bigger!
- arc-gv may generate bigger, more complex CART trees.
 - But they were pruned to 16 leaves.



This Talk

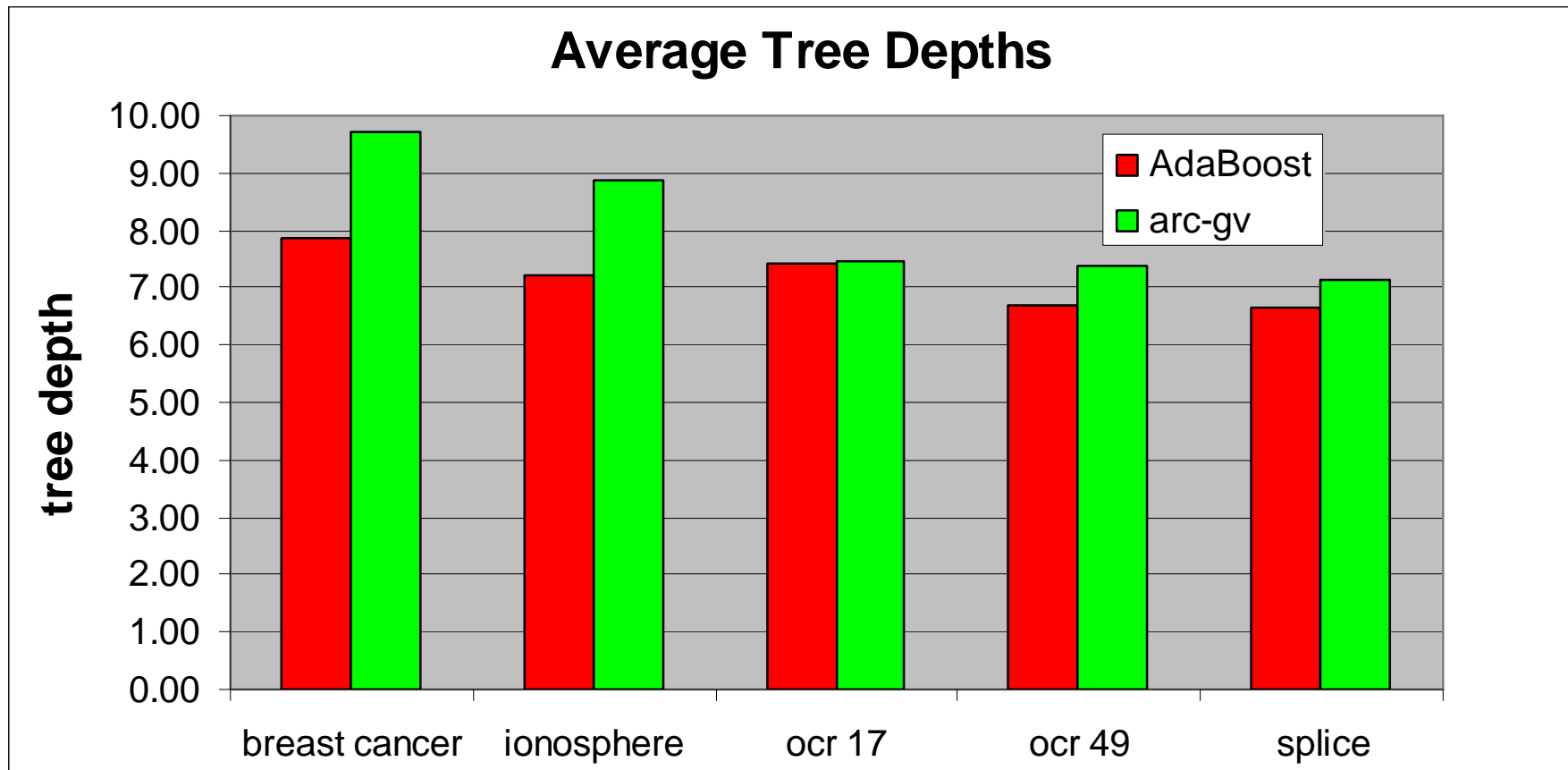
- Take a close look at Breiman's experiments
- Attempt to reconcile with the margins theory by closely examining complexity of the trees

Another Look at the Margins Bound

$$\hat{\Pr} \left[\text{margin}_f(x, y) \leq \theta \right] + \tilde{O} \left(\sqrt{\frac{d}{m\theta^2}} \right)$$


- Maybe tree size is too crude a measure of complexity
- Idea: use tree depth as complexity measure
[Mason et. al. '02]

Measuring Tree Depth



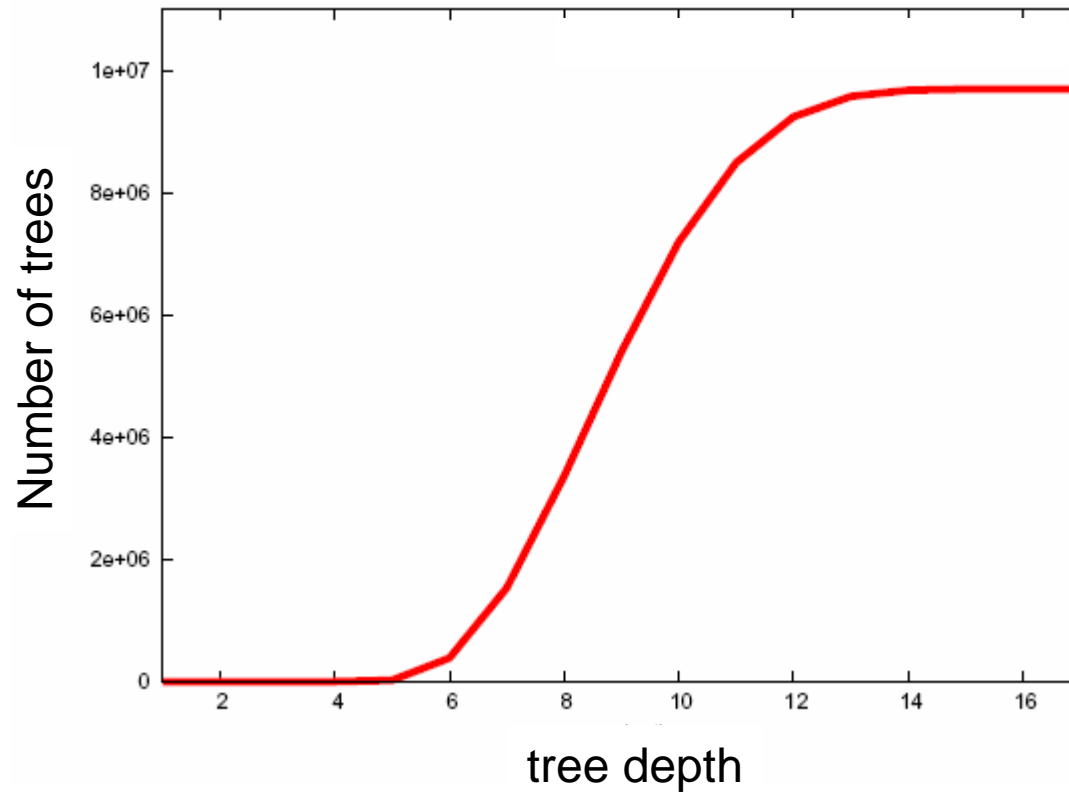


Counting Trees

- We can upper bound the VC-dimension of a finite space of base classifiers H by $\lg |H|$.
- Measuring complexity is essentially a matter of counting how many trees there are of bounded depth.

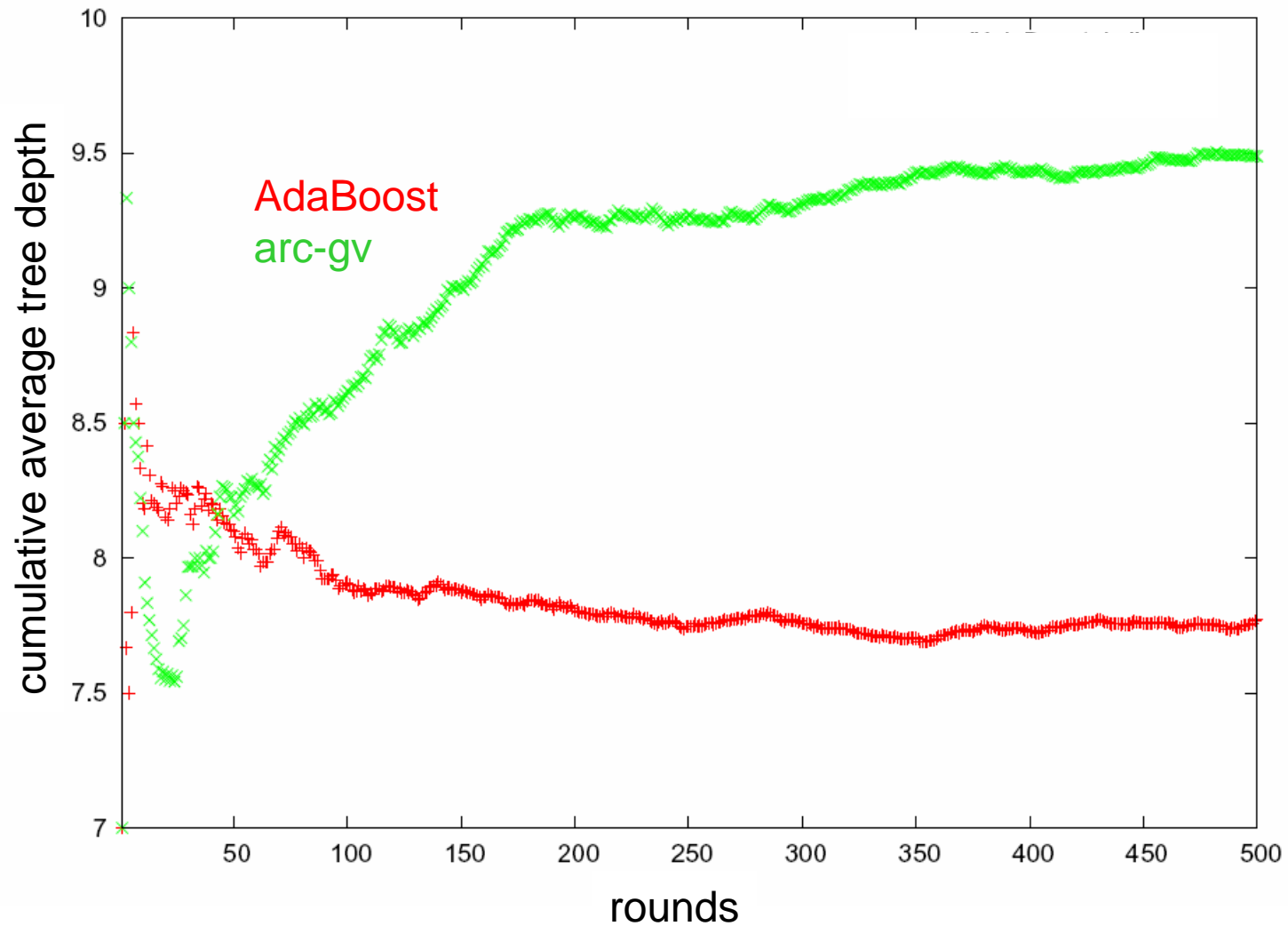


Trees of Bounded Depth



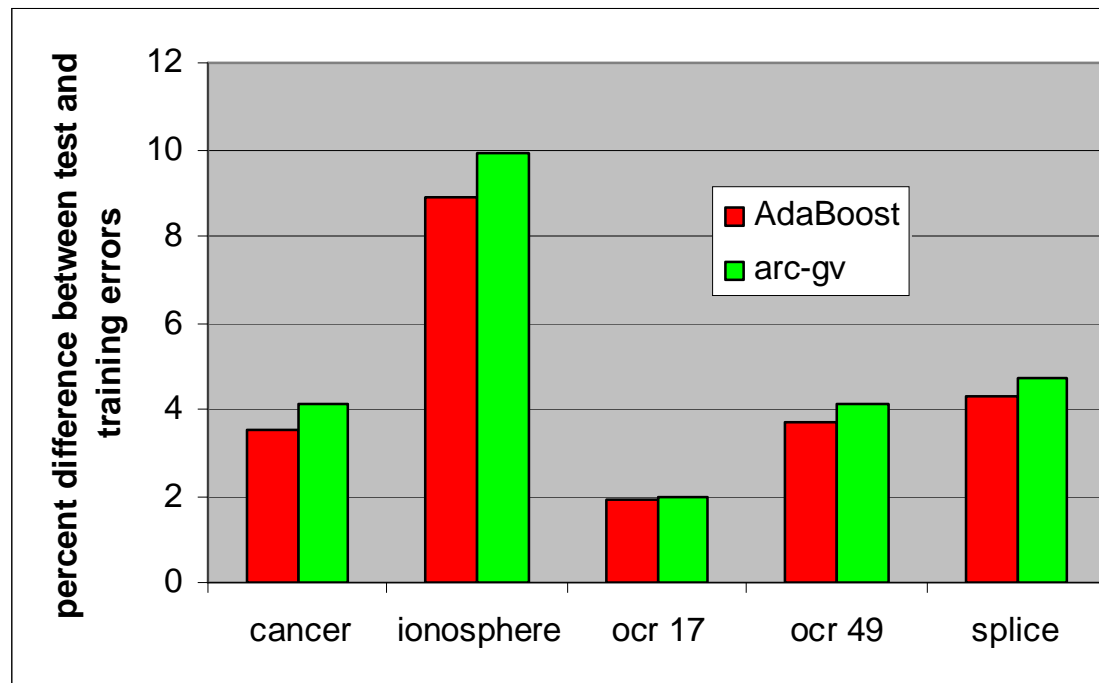


Tree Depth vs Number of Rounds




Another Measure of Tree Complexity

- Idea: difference between training and test error tends to be bigger for higher complexity classifiers.



differences of test and training errors per generated tree averaged over all CART trees in 500 rounds of boosting (over 10 trials)



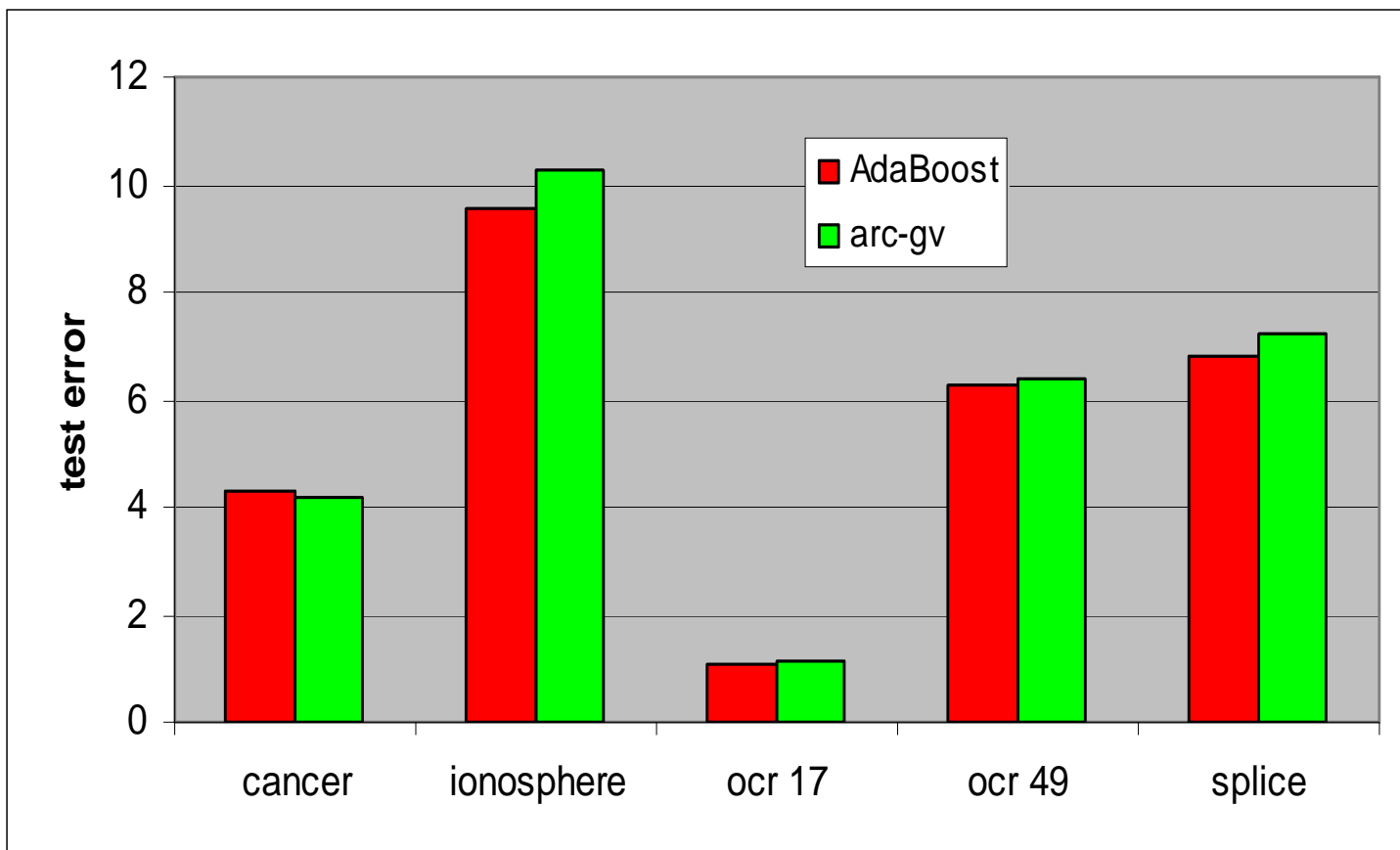
The margins explanation basically says that when all other factors are equal, higher margins result in lower error.

Given that arc-gv tends to choose trees of higher complexity, its higher test error no longer qualitatively contradicts the margin theory.

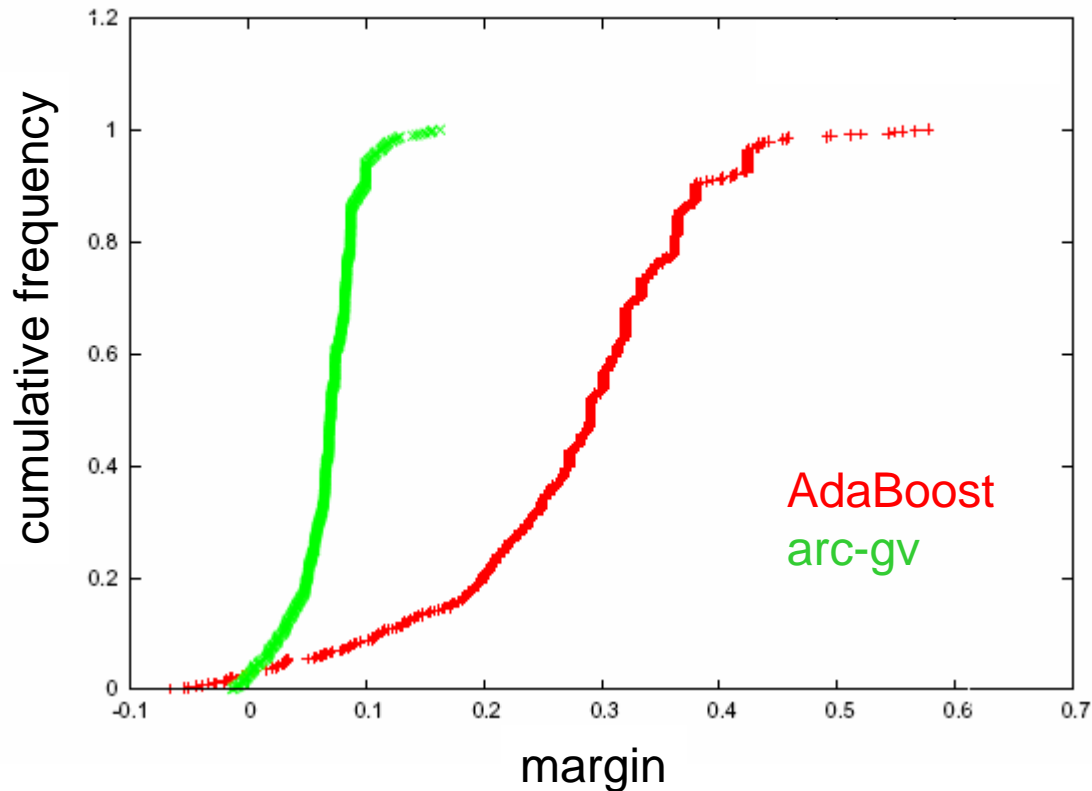
What if we control complexity?

Let's try decision stumps as base classifiers

Controlling Classifier Complexity: Decision Stumps



Decision Stumps



The **minimum** margin is bigger for arc-gv, but the **overall** margins distribution is higher for AdaBoost



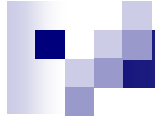
Discussion

- Breiman's results may not actually contradict the margins theory.
- **Margins** are important, but not always at the expense of **other factors**.
- Slightly different boosting algorithms can cause radically different behavior in their generated base classifiers.



Open Questions

- So far, unable to find weak learner of constant complexity with uniformly greater margins distribution for arc-gv than AdaBoost. Does one exist?
- Can we design better boosting algorithms – maximizing average margin?



The End

Thank You