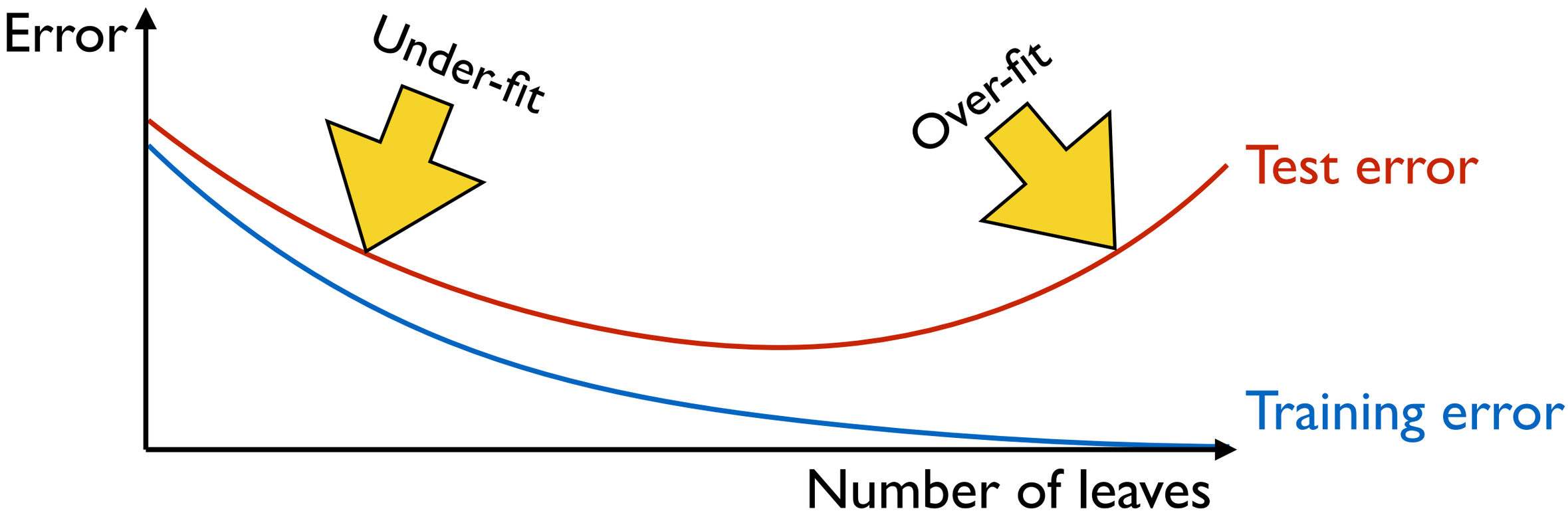
Over-fitting, Ensambles And Margins

The model selection problem

An example with trees

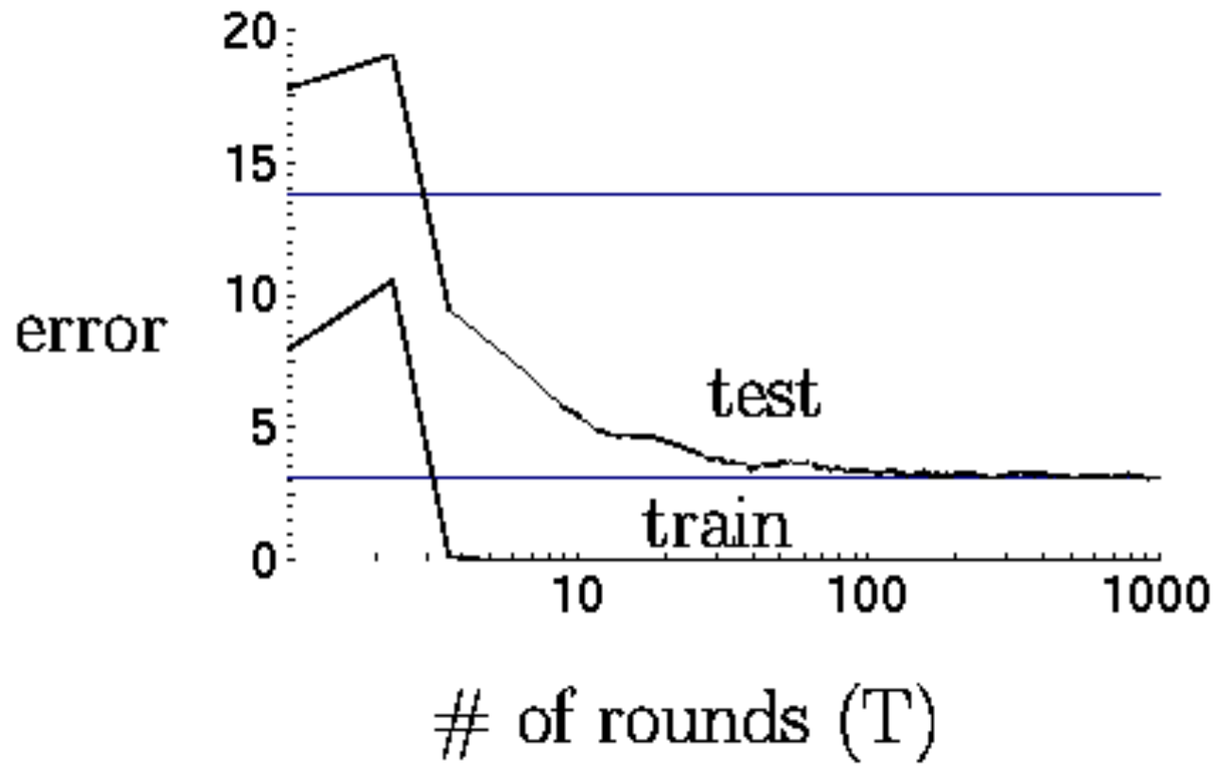
- Train a full decision tree
- Prune tree to have k leaves
- Test tree on test data.



Boosting and over-fitting

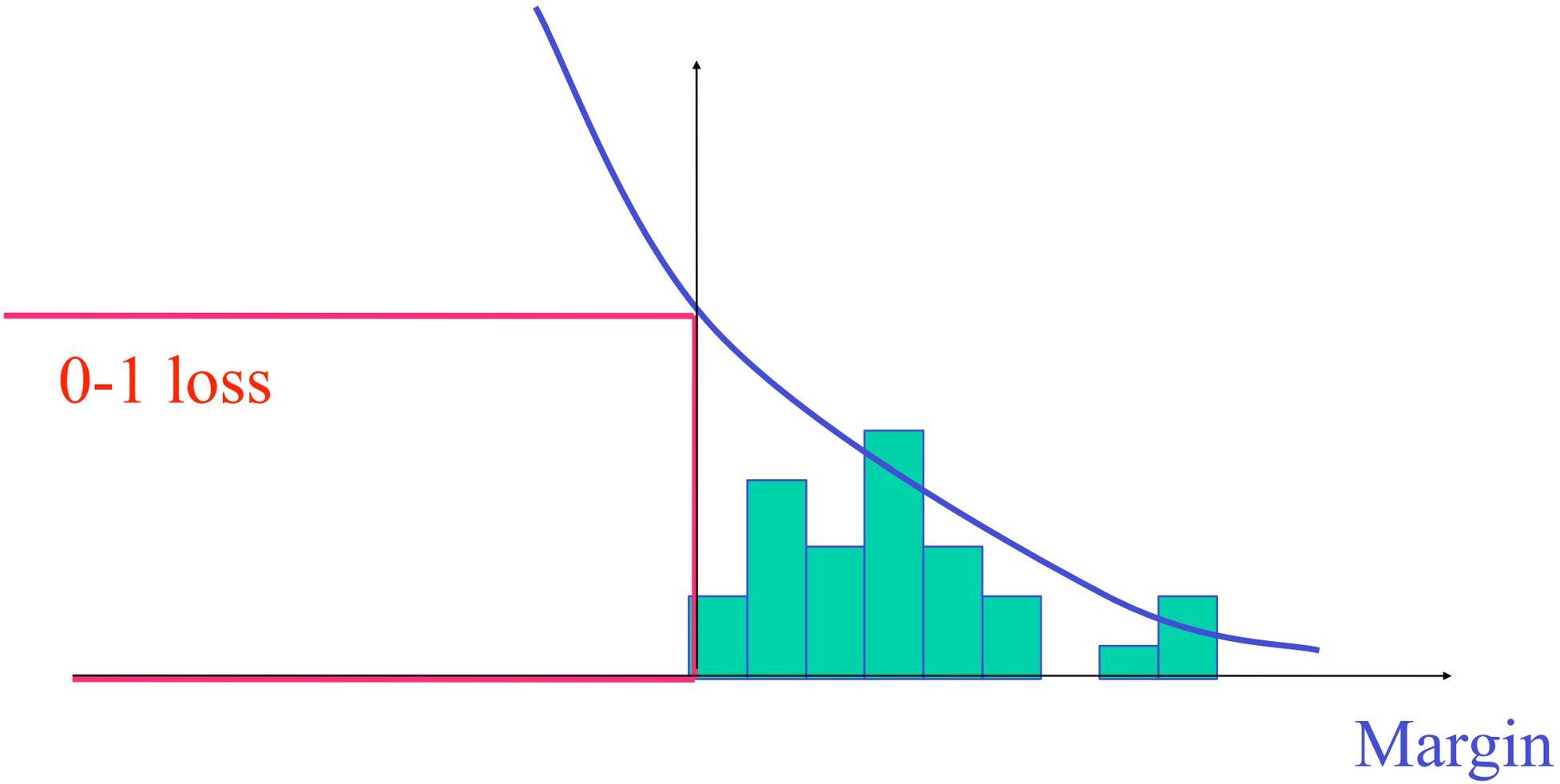
Curious phenomenon

Boosting decision trees

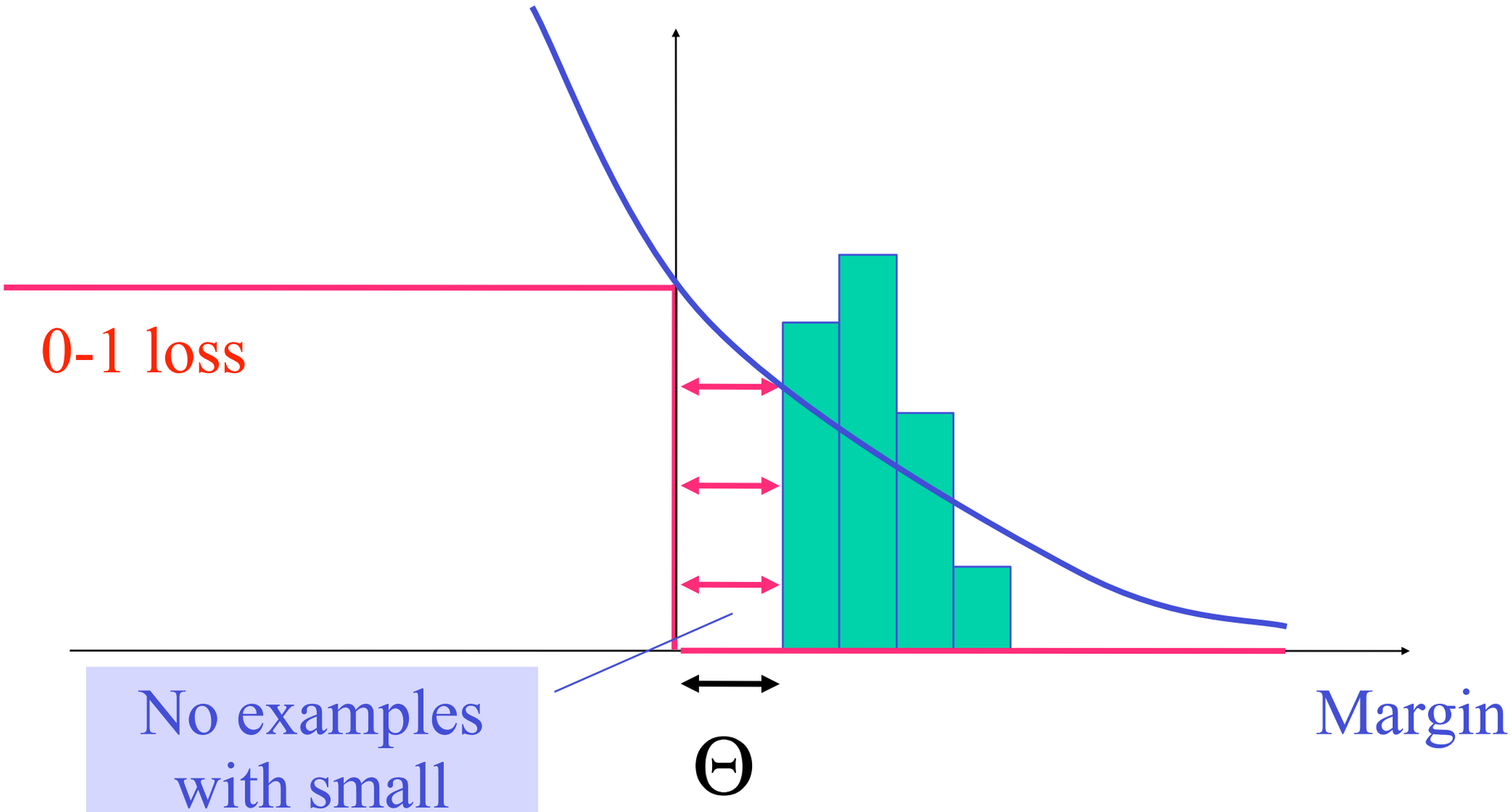


Using <10,000 training examples we fit >2,000,000 parameters

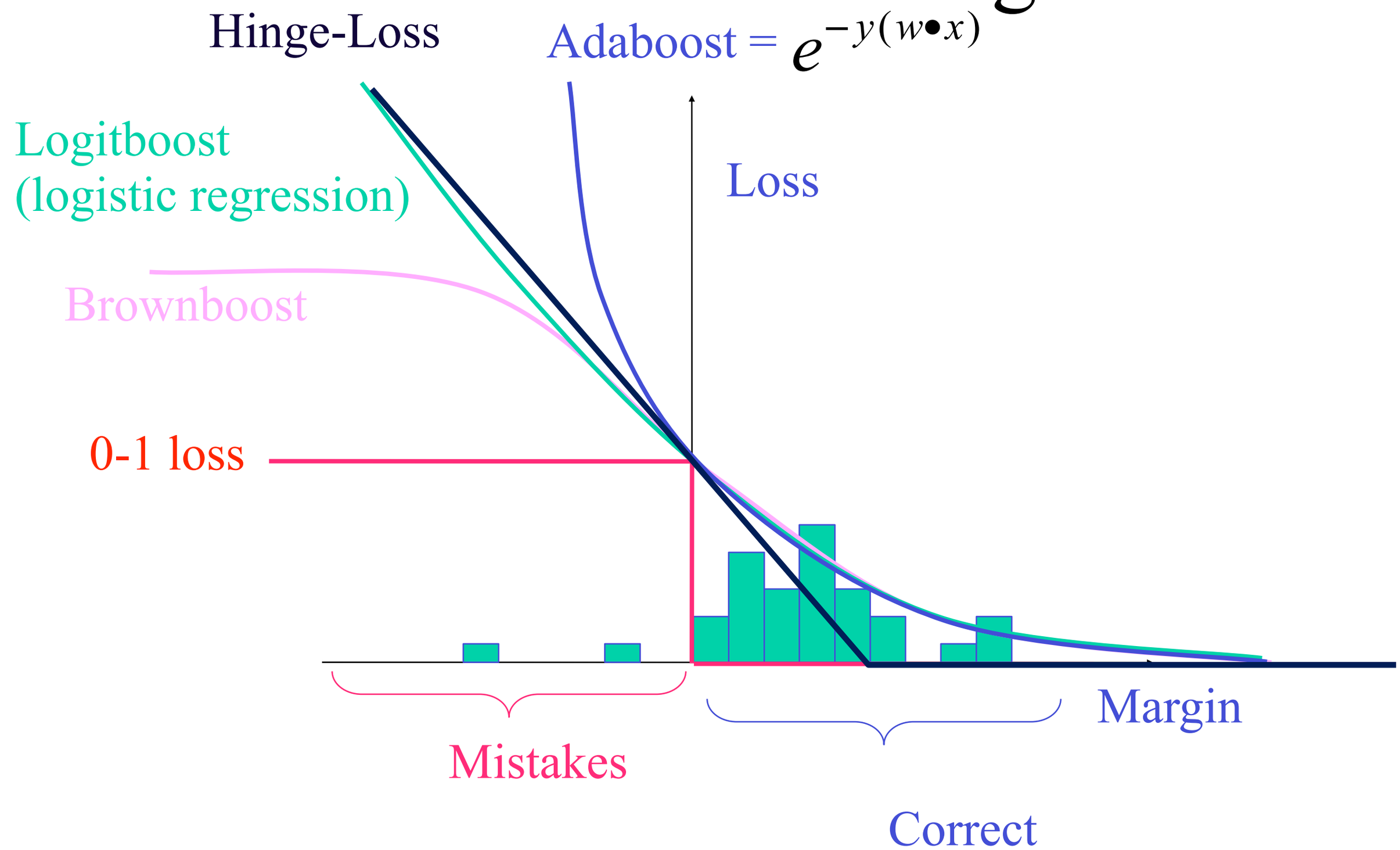
Explanation using margins



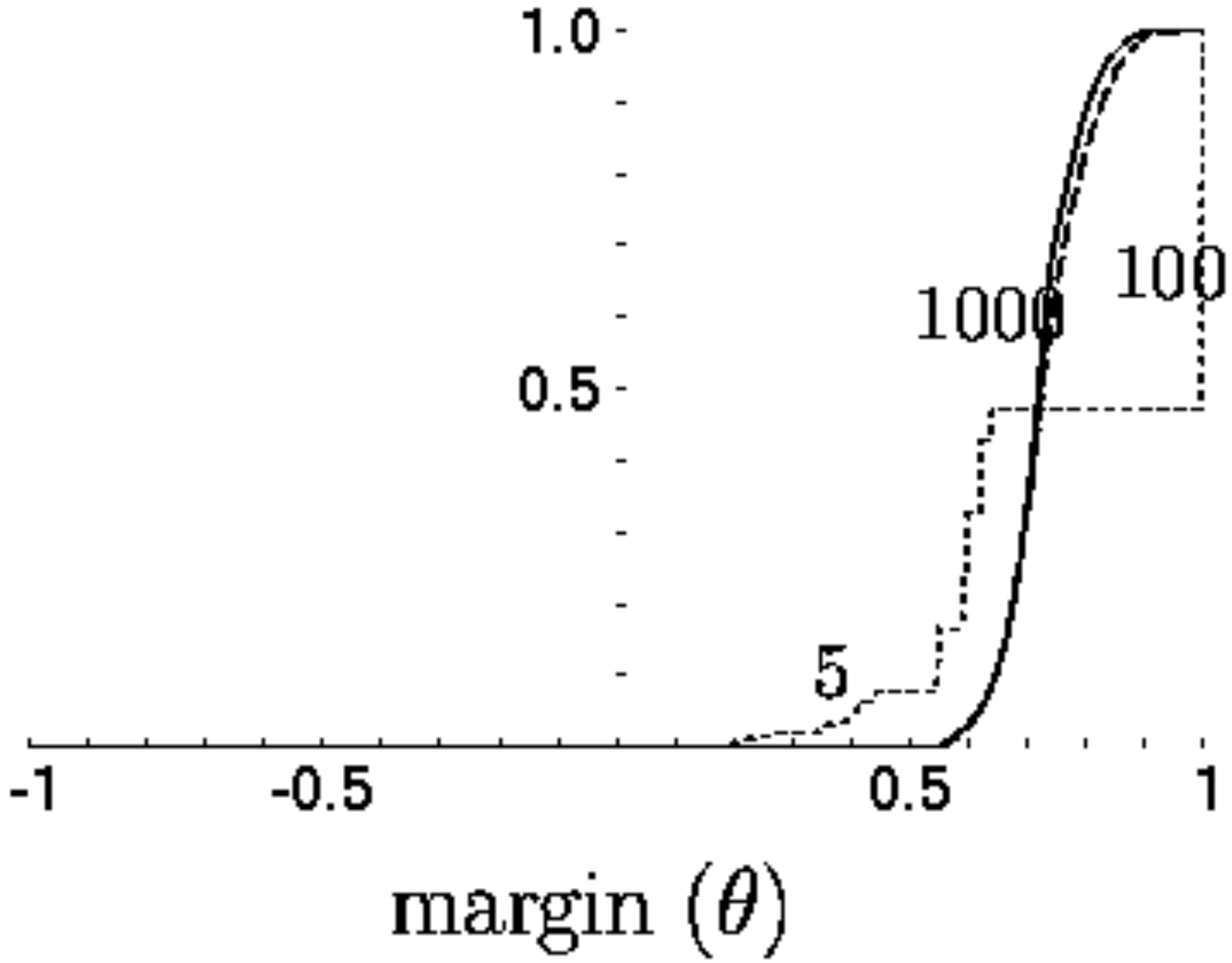
Explanation using margins



Minimizing error using loss functions on margin



Experimental Evidence



Theorem

Schapire, Freund, Bartlett & Lee
Annals of stat. 98

For any convex combination and any threshold $\forall f \in C, \forall \theta > 0$:

Probability of mistake

Fraction of training example with small margin

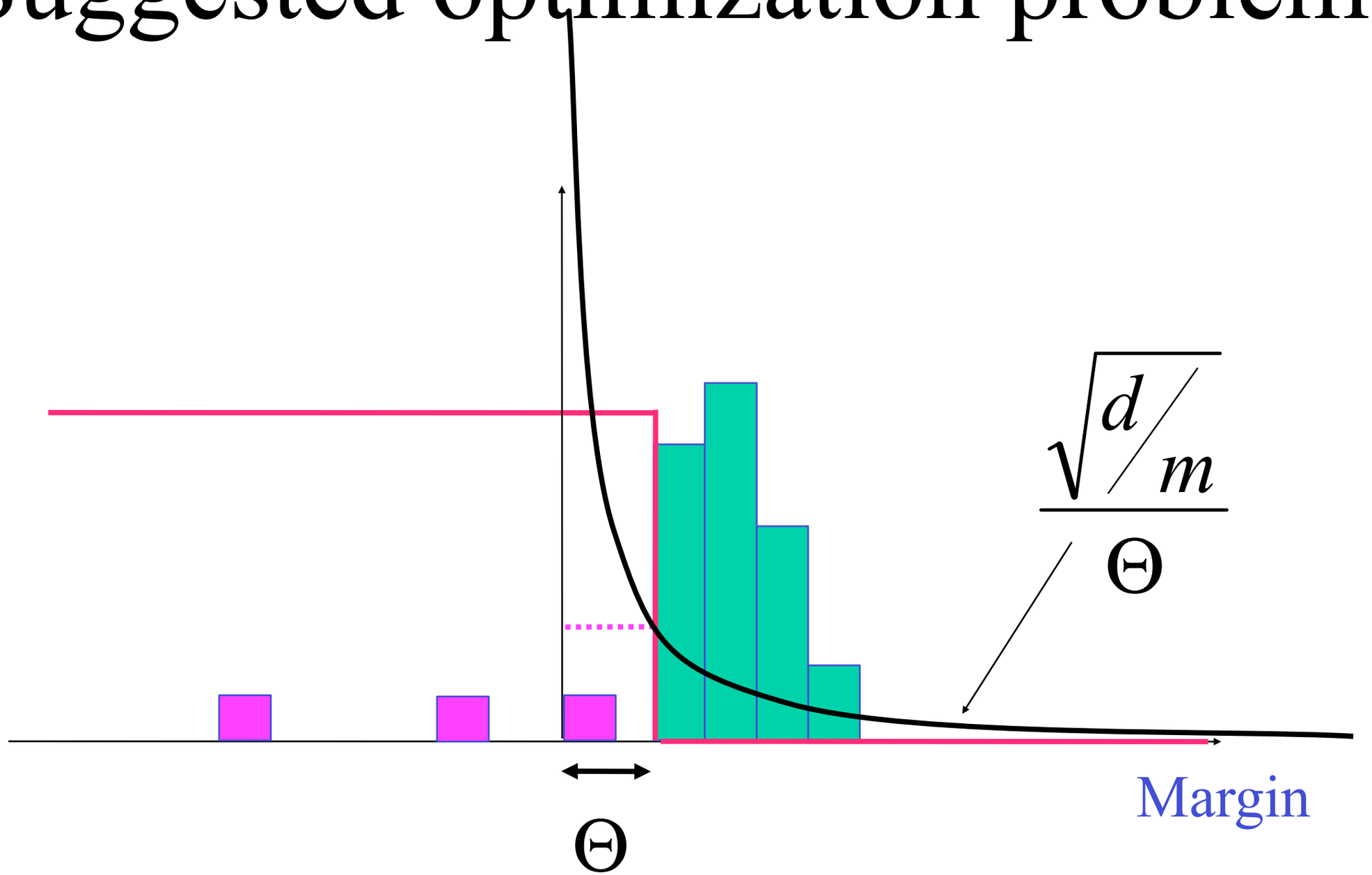
$$P_{(x,y)\sim D}\left[\text{sign}(f(x)) \neq y\right] \leq P_{(x,y)\sim S}\left[margin_f(x,y) \leq \theta\right] + \tilde{O}\left(\frac{\sqrt{d/m}}{\theta}\right)$$

Size of training sample

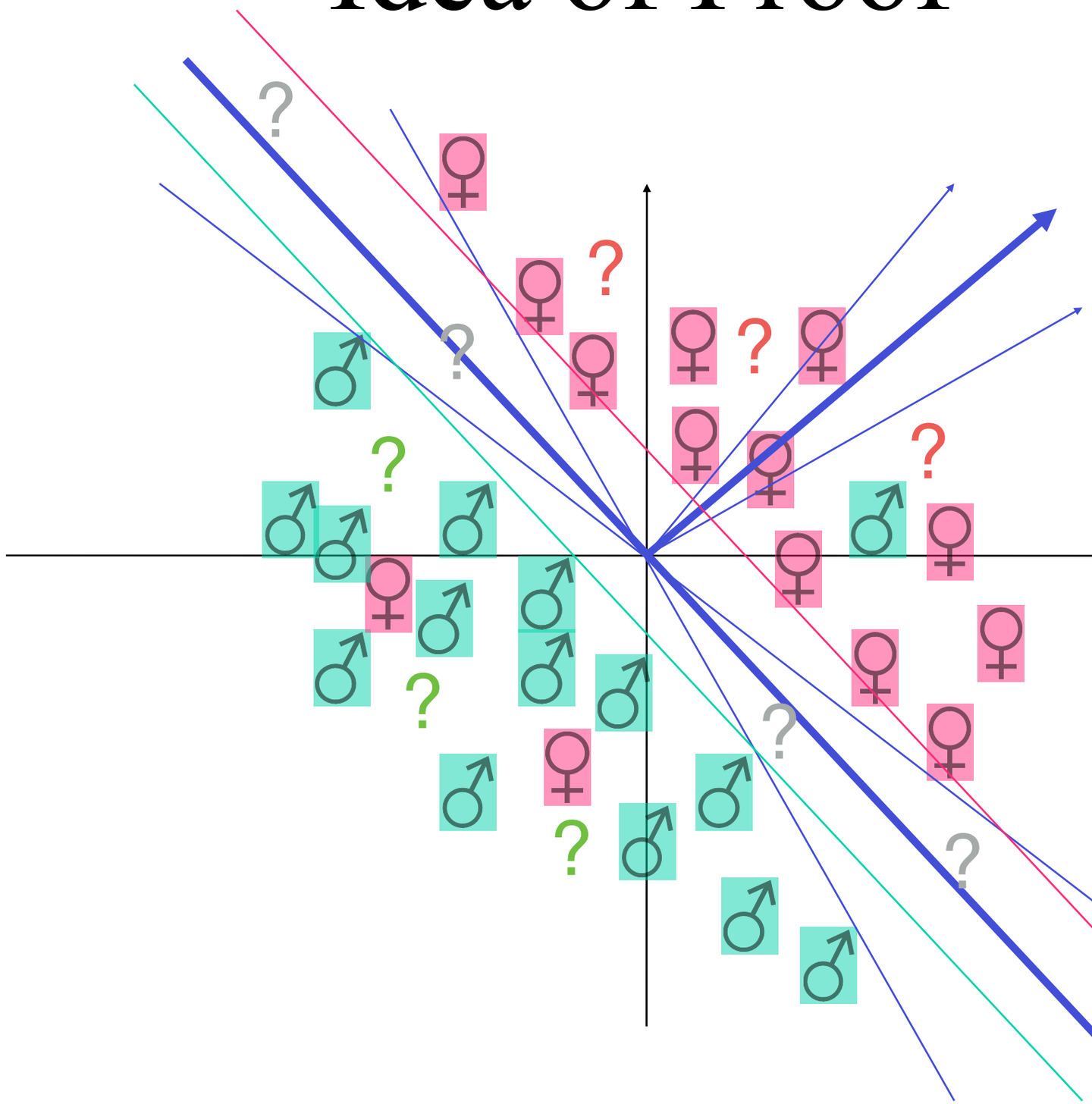
VC dimension of weak rules

No dependence on number of weak rules that are combined!!!

Suggested optimization problem



Idea of Proof



- Find linear classifier with large margin
- Large margin implies there is a **cone** of linear classifiers with close-to-minimal error.
- Large margin test examples are classified identically by classifiers in cone.
- The prediction on large margin test examples is insensitive to small changes in training set.

The three sources of classification error

- **Bayes Error:** the minimal error rate achieved when the true distribution is known (exactly).
- **Model Bias:** error resulting from the inability of the model class to represent the true distribution. (Under-fit)
- **Data Variation:** error resulting from the difference between the training set and the true distribution. (Over-fit)

Example, Bayes Error

- X={1,2,3} dist. is uniform.(1/3,1/3,1/3)
- Y={-1,+1},
  - P(y=+1|x=1)=0.1
  - P(y=+1|x=2)=0.8
  - P(y=+1|x=3)=0.3
- Bayes optimal classifier: f(1)=-1,f(2)=+1,f(3)=?
  - f(3)=-1
- Bayes Error= (0.1+0.2+?)/3
  - (0.1+0.2+0.3)/3=0.6/3 = 0.2

Example, Model Bias

- Bayes optimal classifier: f(1)=-1,f(2)=+1,f(3)=-1

- Suppose our model only allows rules of the form

$$g_\theta(x) = \begin{cases} +1 & \text{if } x > \theta \\ -1 & \text{otherwise} \end{cases}$$

- Cannot represent Bayes optimal classifier
- best threshold: $\theta = 1.1$ (or any number between 1 and 2)
- Error of best threshold
  - (0.1+0.2+0.7)/3=1/3=0.333
  - Compare with Bayes:(0.1+0.2+0.3)/3=0.2

Example, Data Variation (2)

- In practice, we don't know the conditional probabilities, we estimate them from data:

$* x=1,\ 1\times(y=+1), 6\times(y=-1) \Rightarrow \hat{P}(y=+1\,|\,x=1)=1/7=0.14 \quad P(y=+1\,|\,x=1)=0.1$

$* x=2,\ 8\times(y=+1), 2\times(y=-1) \Rightarrow \hat{P}(y=+1\,|\,x=2)=0.8 \quad P(y=+1\,|\,x=2)=0.8$

$* x=3,\ 2\times(y=+1), 3\times(y=-1) \Rightarrow \hat{P}(y=+1\,|\,x=3)=0.4 \quad P(y=+1\,|\,x=3)=0.6$

- For x=3 estimate is below 0.5, truth is above 0.5
- Empirically optimal classifier : f(1)=-1, f(2)=+1, f(3)=-1
- This rule is NOT the Bayes optimal rule, it was misled by the empirical estimate for x=3. We say that the discrepancy is due to data variation.
- The best rule of the form $g_\theta(x)=\begin{cases} +1 & \text{if } x>\theta \\ -1 & \text{otherwise} \end{cases}$ is equal to the Bayes optimal rule.
- **<u>Restricting the model class increases model bias while reducing data variation.</u>**

Methods for reducing data variation for classifiers

- **Model Selection:** Choosing the number of parameters in the model (number of leaves in a tree).
- **Regularization:** Adding a term that penalizes for size of weights (regression, neural networks)
- **Margins:** increasing the classification margin. (Boosting, Support Vector Machines)
- **Bagging:** Taking the majority vote over randomize choice of training set.

Tradeoffs

- Weight decay, reducing the number of parameters decrease data variation, increase model bias.
- Bagging decreases data variation without significantly increasing model bias.
- Boosting reduces both variation and model bias.

The Bias/Variance tradeoff for **squared error**

- The squared error can always be decomposed into a bias term and a variance term.

$B(x) = E_{(x,y)}[y \mid x]$ The Bayes optimal regression function

$$E_{(x,y)}\left[\left(f(x)-y\right)^2\right] = \overbrace{E_x\left[\left(f(x)-B(x)\right)^2\right]}^{Bias} + \overbrace{E_{(x,y)}\left[\left(y-B(x)\right)^2\right]}^{Variance}$$

- For regression:
  - Increasing the number of features decreases bias and increases variance.
  - Reducing the size of the weights (Ridge regression, Lasso) Increases bias and decreases variance.

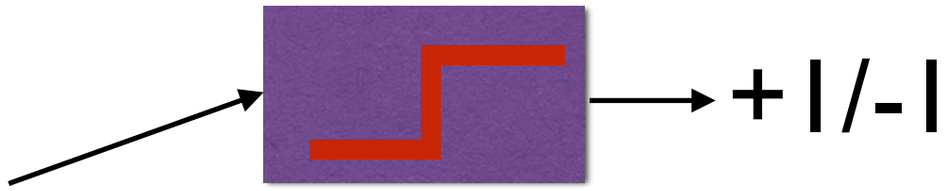
Why are large margins good?

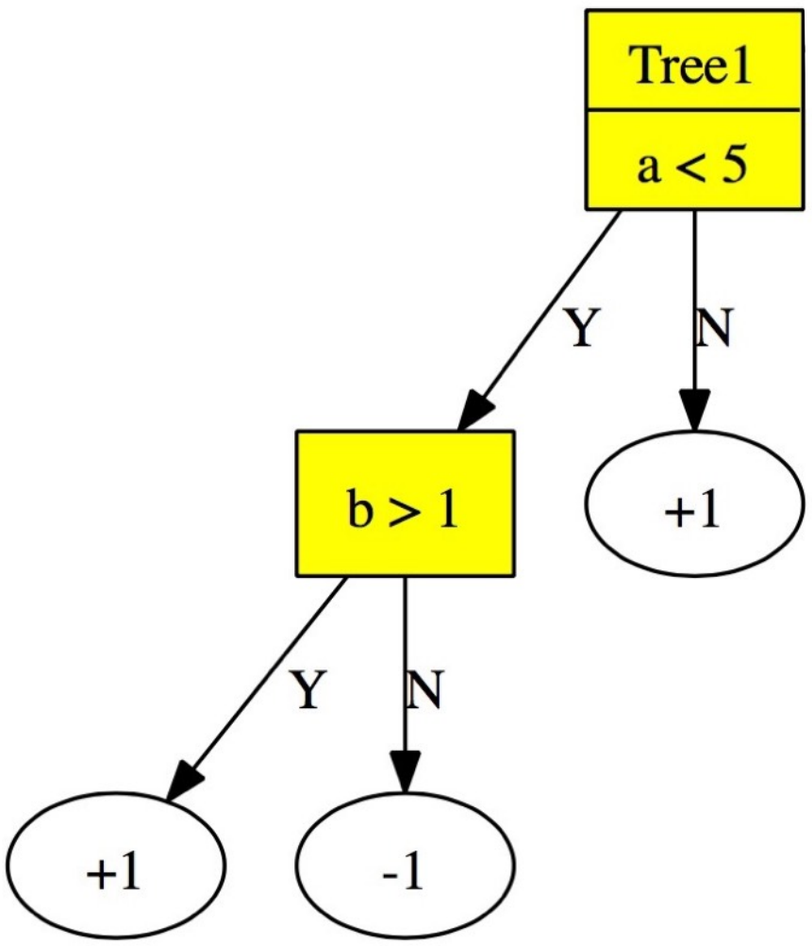
- On examples with large margin, the prediction is stable.
- In other words, small changes in the training set will not cause the prediction to flip.
- We can predict with confidence.
- The confidence is different from when P(y|x) is close to 0 or to 1.
- It has to do with the **stability** of the classifier w.r.t. small changes in the training data.

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.
- Any predictor can be used as a base predictor.
  - In this talk,
  - We will restrict our attention to binary classification, but there are solutions for multi class and for regression.

An Ensemble of trees





The Bootstrap

1990

An Introduction to the Bootstrap

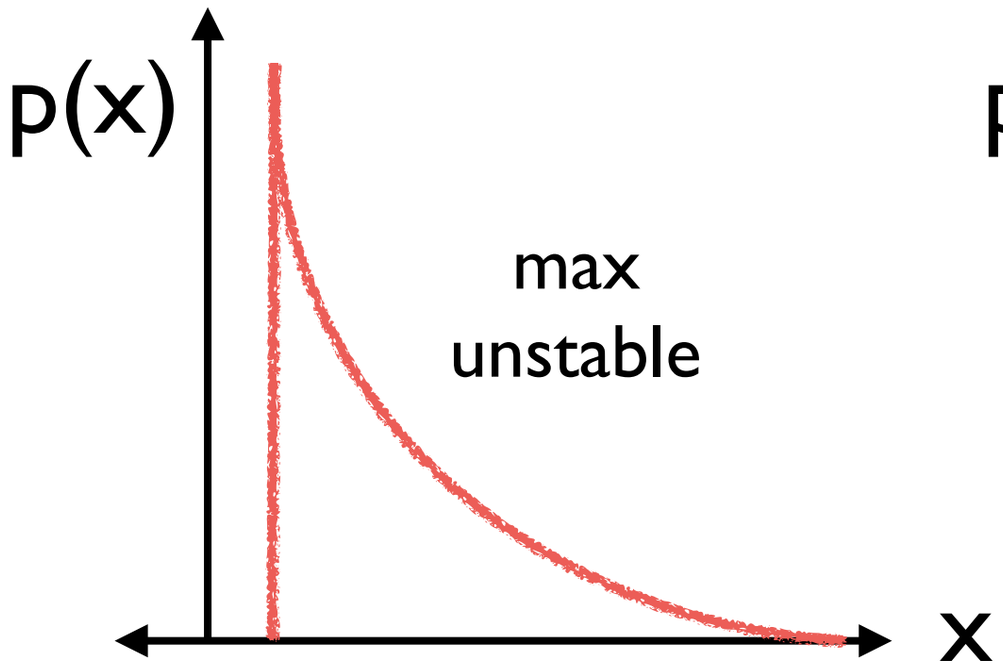
Bradley Efron
Department of Statistics
Stanford University
and
Robert J. Tibshirani
Department of Preventative Medicine and Biostatistics
and Department of Statistics, University of Toronto

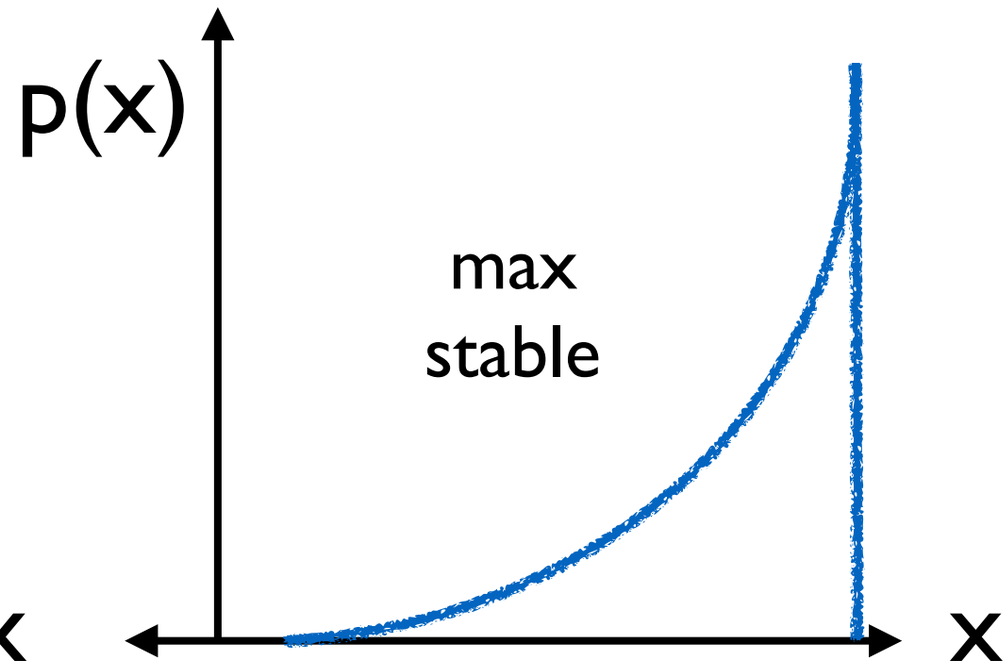
CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

The Bootstrap

- A method for estimating out-of-sample variation
- Suppose that we estimate the maximum of of the distribution.
- Depending on the distribution, this estimate of max might be stable or not.





How to estimate the variation

- In general - a hard question, we are trying to estimate a property of the true distribution, but we only have the sample.
- A Bootstrap sample: given a sample of size n, sample n times with replacement from this sample.
- Compute the estimator on each bootstrap sample an see how much it varies.
- Will work nicely for the max estimator.

Bagging = bootstrap aggregation

- Decision trees have high data variation.
    - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.
    - Running an algorithm over random resampling is called “The Bootstrap”
- Trees can be learned in parallel
- The result is a reduction in variation with no increase in the bias.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.
- Speeds up the learning process.
- All trees get equal weight (1.0)
- All trees can be learned in parallel.

Random Forests and Bagging

- Bagging: Bootstrap Aggregation
- Bagging+Decision threes = Random forests
- Run CART, considering a random small subset of the features at each iteration.
- Examples are NOT reweighted.
- Take a majority vote over many trees.
- A very fast, flexible and accurate algorithm.
- Natural notion of margins.
- Faster than boosting and as accurate.
- The generated rules are bigger than corresponding ADTrees.

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.
- Each tree is trained using a **weighted** training set. The weights represent the gradient of the loss function.
- Each tree receives a different weight (corresponding to the alpha in adaboost)
- Stochastic gradient boosting: use random resampling of the training set a.k.a. Bagging.