Many effective ensemble methods

- Sequential methods
  - AdaBoost [Freund & Schapire, JCSS97]
  - Arc-x4 [Breiman, AnnStat98]
  - LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
  - ... ...
- Parallel methods
  - Bagging [Breiman, MLJ96]
  - Random Subspace [Ho, TPAMI98]
  - Random Forests [Breiman, MLJ01]
  - ... ...

Generalization bound

Freund & Schapire [JCSS97] proved that the generalization error of AdaBoost is bounded by:

$$\epsilon_{\mathcal{D}} \leq \epsilon_D + \tilde{O}\left(\sqrt{\frac{dT}{m}}\right)$$

with probability at least $1-\delta$, where d is the **VC-dimension** of base learners, m is the number of training instances, T is the number of learning rounds and $\tilde{O}(\cdot)$ is used instead of $O(\cdot)$ to hide logarithmic terms and constant factors.

Generalization bound

Freund & Schapire [JCSS97] proved that the generalization error of AdaBoost is bounded by:

$$\epsilon_{\mathcal{D}} \leq \epsilon_D + \tilde{O}\left(\sqrt{\frac{dT}{m}}\right)$$

with probability at least $1-\delta$, where d is the **VC-dimension** of base learners, m is the number of training instances, T is the number of learning rounds and $\tilde{O}(\cdot)$ is used instead of $O(\cdot)$ to hide logarithmic terms and constant factors.

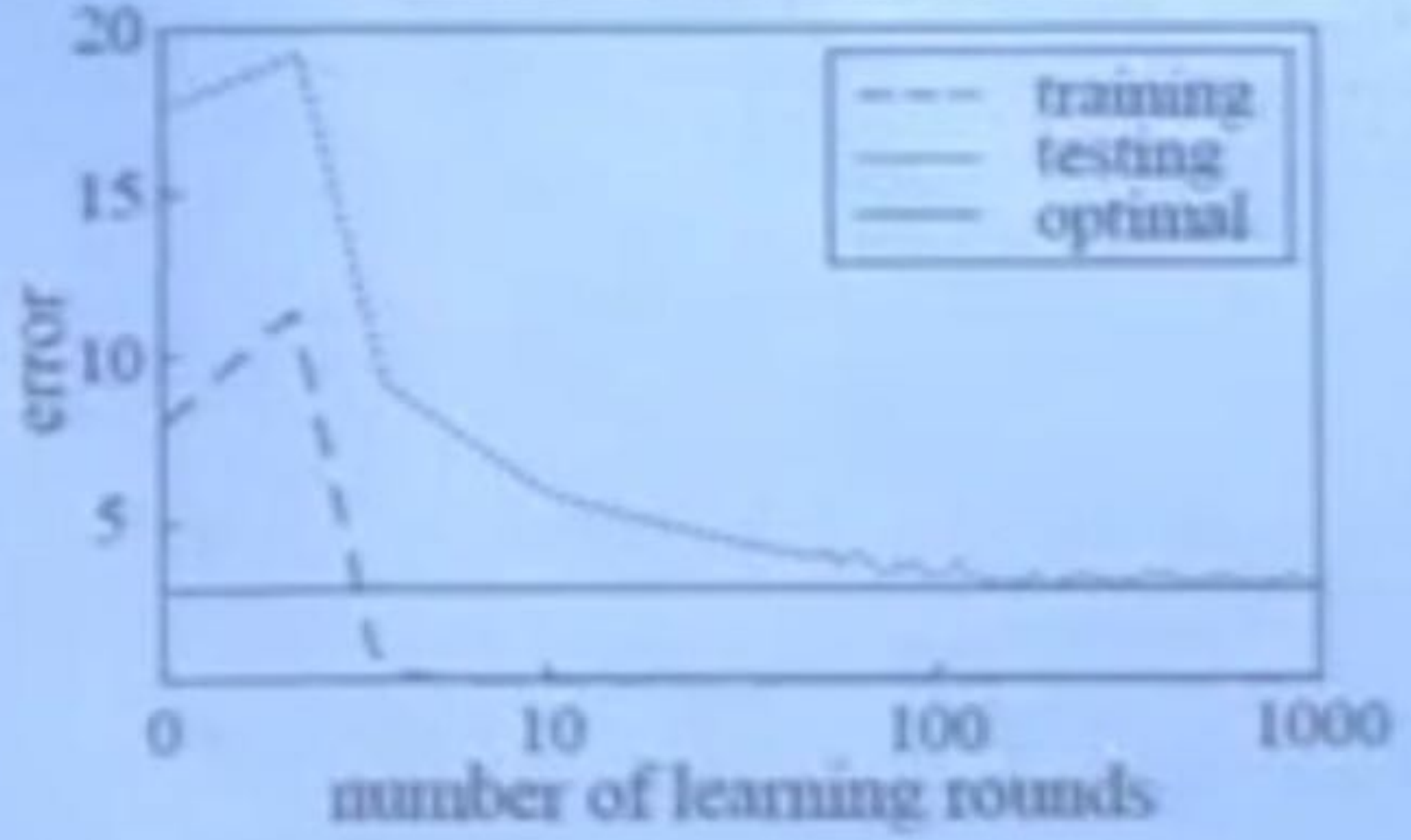
It implies that AdaBoost will **overfit** if T is large

Overfit (过拟合): The trained model fits the training data too much such that it can exaggerate minor fluctuations in the training data, leading to poor generalization performance

The Mystery

However, AdaBoost often does not overfit in real practice

A typical performance plot of AdaBoost on real data



Seems contradict with the **Occam's Razor**

Knowing the reason may inspire new methodology for algorithm design

Major theoretical efforts

- **Margin Theory**

  Started from [Schapire, Freund, Bartlett & Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26(5):1651–1686, 1998]

- **Statistical View**

  Started from [Friedman, Hastie & Tibshirani. Additive logistic regression: A statistical view of boosting (with discussions). Annals of Statistics, 28(2):337–407, 2000]

Major theoretical efforts

- **Margin Theory**

  Started from [Schapire, Freund, Bartlett & Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26(5):1651–1686, 1998]

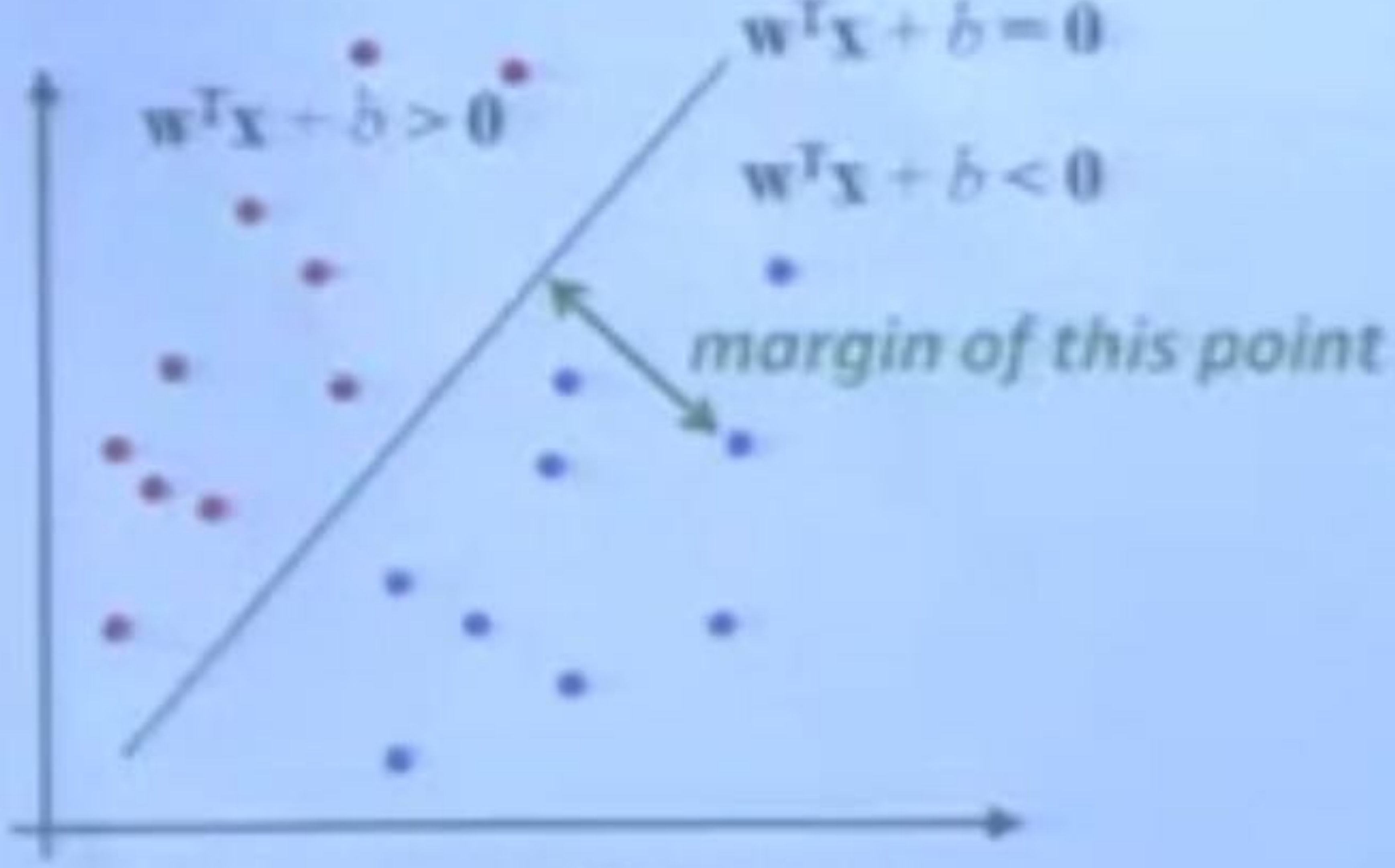
- Statistical View

  Started from [Friedman, Hastie & Tibshirani. Additive logistic regression: A statistical view of boosting (with discussions). Annals of Stati

The biggest issue:
The statistical view did not explain why AdaBoost is resistant to overfitting

The "margin" (间隔)

Binary classification can be viewed as the task of separating classes in a feature space



$f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$

The bigger the margin,
the higher the predictive confidence

For binary classification, the ground-truth $f(\boldsymbol{x}) \in \{-1, +1\}$

The margin of a single classifier h: $f(\boldsymbol{x})h(\boldsymbol{x})$

For $H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}_t)$
the margin is

$$f(\boldsymbol{x})H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})$$

and the normalized margin:

$$\frac{\sum_{t=1}^{T} \alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})}{\sum_{t=1}^{T} \alpha_t}$$

Margin explanation of AdaBoost

Based on the concept of margin, Schapire et al. [1998] proved that, given any threshold $\theta > 0$ of margin over the training data D, with probability at least $1-\delta$, the generalization error of the ensemble $\epsilon_{\mathcal{D}} = P_{\boldsymbol{x}\sim\mathcal{D}}(f(\boldsymbol{x}) \neq H(\boldsymbol{x}))$ is bounded by

$$\begin{aligned}\epsilon_{\mathcal{D}} &\leq P_{\boldsymbol{x}\sim D}(f(\boldsymbol{x})H(\boldsymbol{x}) \leq \theta) + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2} + \ln\frac{1}{\delta}}\right)\\ &\leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2} + \ln\frac{1}{\delta}}\right)\end{aligned}$$

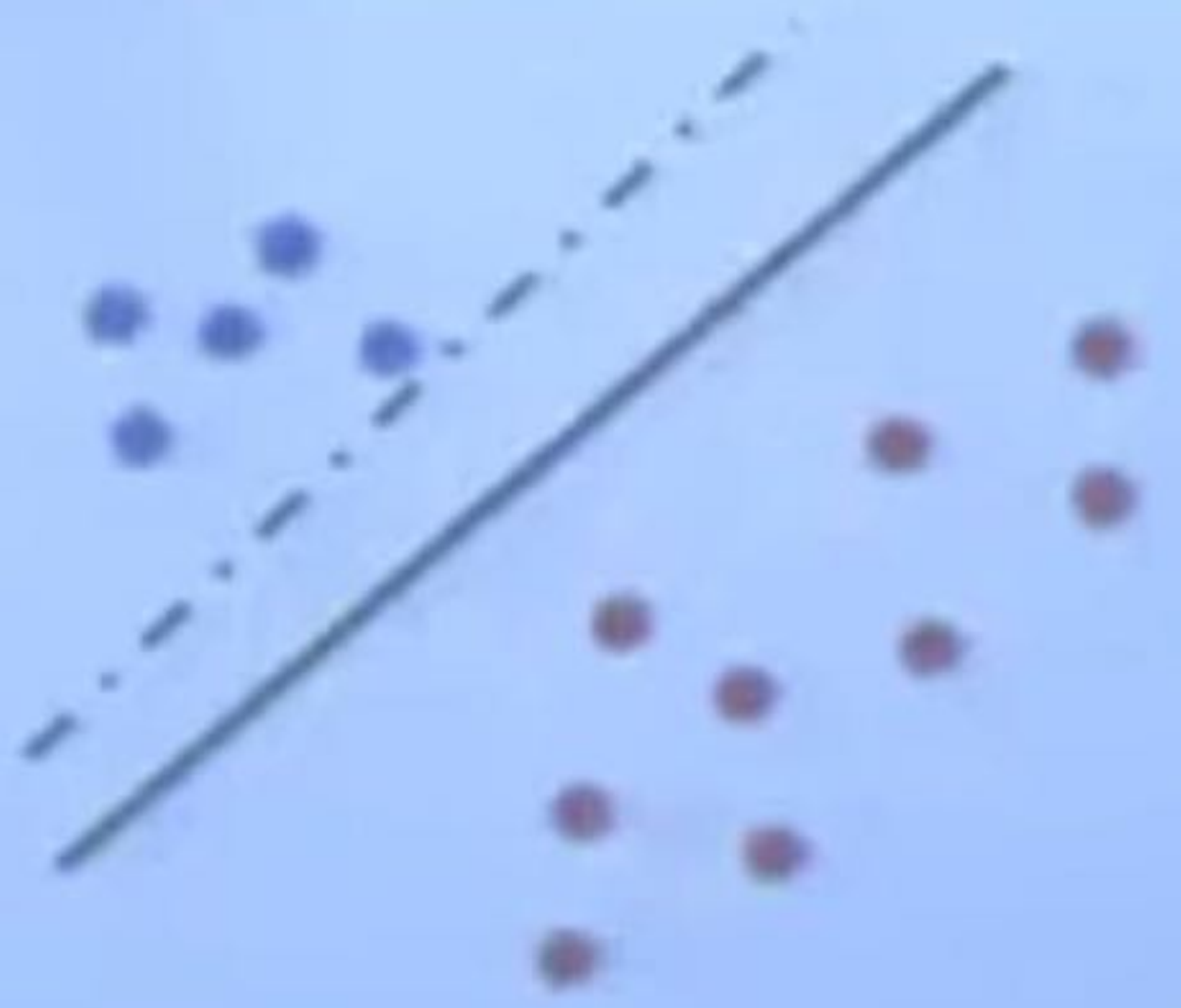
This bound implies that, when other variables are fixed, the larger the margin over the training data, the smaller the generalization error

Margin explanation of AdaBoost (con't)

Why AdaBoost tends to be resistant to overfitting?

the margin theory answers:

Because it is able to increase the ensemble margin even after the training error reaches zero



This explanation is quite intuitive

It receives good support in empirical study

The minimum margin bound

Schapire et al.'s bound depends heavily on the smallest margin, because $P_{x\sim D}(f(x)H(x) \leq \theta)$ will be small if the smallest margin is large

Thus, by considering the minimum margin:

$$\varrho = \min_{x \in D} f(x)H(x)$$

Breiman [Neural Comp. 1999] proved a generalization bound, which is tighter than Schapire et al.'s bound

The doubt about margin theory

Breiman [Neural Comp. 1999] designed a variant of AdaBoost, the arc-gv algorithm, which directly maximizes the minimum margin

> the margin theory would appear to predict that arc-gv should perform better than AdaBoost

However, experiments show that, comparing with AdaBoost:

- arc-gv does produce uniformly larger minimum margin
- the test error increases drastically in almost every case

The doubt about margin theory

Breiman [Neural Comp. 1999] designed a variant of AdaBoost, the arc-gv algorithm, which directly maximizes the minimum margin

the margin theory would appear to predict that arc-gv should perform better than AdaBoost

However, experiments show that, comparing with AdaBoost:

- arc-gv does produce uniformly larger minimum margin
- the test error increases drastically in almost every case

Thus, Breiman convincingly concluded that **the margin theory was in serious doubt**. This almost sentenced the margin theory to death

Long march of margin theory for AdaBoost

- 1989, [Kearns & Valiant], open problem
- 1990, [Schapire], proof by construction, the first Boosting algorithm
- 1993, [Freund], another impractical boosting algorithm by voting
- 1995/97, [Freund & Schapire], AdaBoost

- 1998, [Schapire, Freund, Bartlett & Lee], Margin theory
- 1999, [Breiman], serious doubt by minimum margin bound
- 2006, [Reyzin & Schapire], finding the model complexity issue in exps, emphasizing the importance of margin distribution
- 2008, [Wang, Sugiyama, Yang, Zhou & Feng], Emargin bound, believed to be a margin distribution bound
- 2013, [Gao & Zhou], a real margin distribution bound, sheding new insight ; margin theory defensed


Deep models revisited
LAMDA
Learning And Mining from DatA
http://www.lamda.nju.edu.cn
Currently, Deep Models are DNNs:
multiple layers of parameterized
differentiable nonlinear modules that
can be trained by backpropagation
• Not all properties in the world are "differentiable", or best modelled as "differentiable"
• There are many non-differentiable learning modules (not able to be trained by backpropagation)
开启AI，开创未来
Open the AI, Open the Future
南京论坛
NANJING FORUM
bilibili BV1mJ411976u 10:23/27:54