28 March 2023

CIS 5200: MACHINE LEARNING

BOOSTING

Surbhi Goel

Content here draws from material by Rob Schapire (Princeton), Cynthia Rudin (Duke), and Kilian Weinberger (Cornell)

Spring 2023

OUTLINE - TODAY

- History
- Setup
- General Boosting Algorithm
- AdaBoost
    - Example
    - Proof of Convergence
    - Generalization
    - Optimization Viewpoint

RECALL

Last class we studied **Bagging**

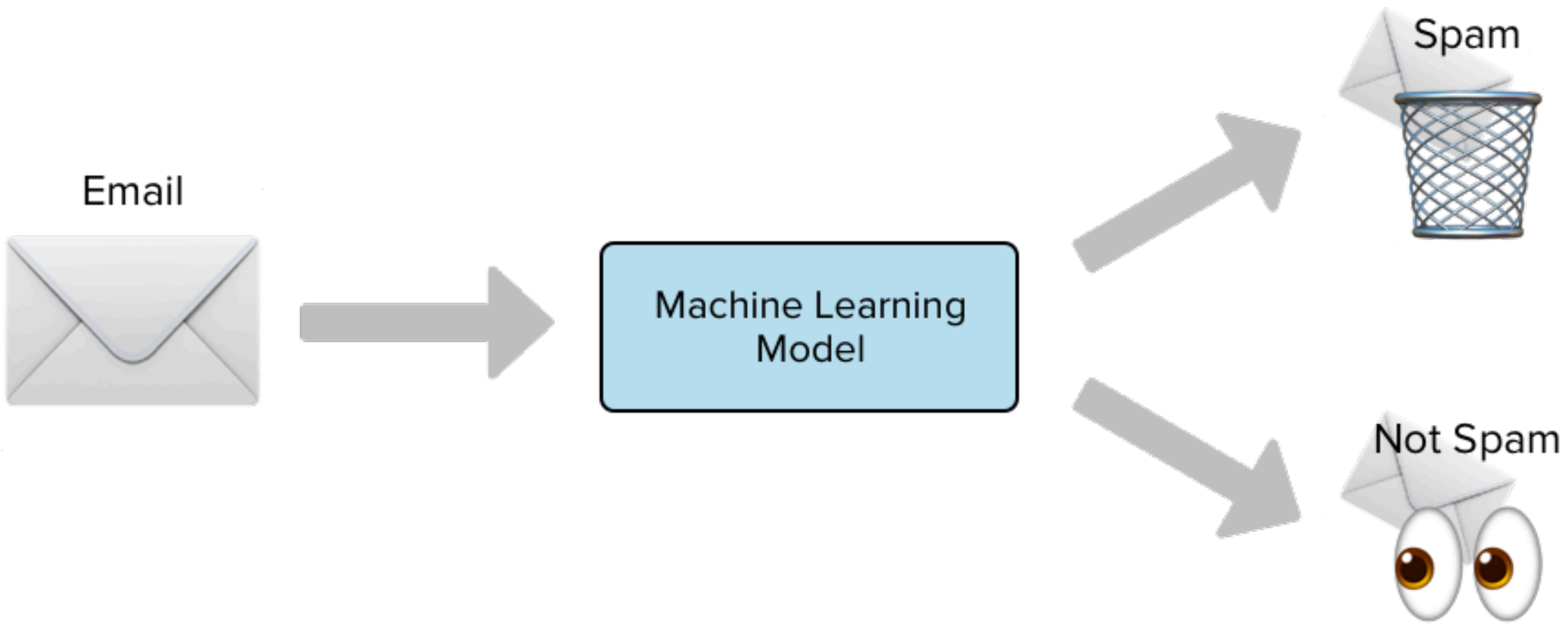
- Generate new training datasets by sampling with replacement from the empirical distribution
- Aggregate learned classifiers on each new training dataset
- Helps reduce variance

Today: How to reduce bias?

SIMPLE LEARNERS

High bias

Suppose you have **simple/weak** learners that are correct $\sim 55\ \%$ of the time (slightly better than random guessing)



Learner 1:
Does the email contain the word 'free'?

Learner 2:
Does the email contain more than half the letters capitalized?

Learner 3:
Does the email contain a long URL address?

CAN WE BOOST SIMPLE LEARNERS?

High bias

Suppose you have **simple/weak** learners that are correct $\sim 55\ \%$ of the time (slightly better than random guessing)

Question: Can we combine a bunch of these **simple/weak** learners to get a **complex/strong** learner that gets close to perfect accuracy?

Famously asked by Michael Kearns in 1988 in a machine learning class project!

YES, WE CAN!

Question (1988): Can we combine a bunch of these **simple/weak** learners to get a **complex/strong** learner that gets close to perfect accuracy?

Answer 1 (1990): Yes, we can! Weak learnability implies strong learnability, and vice-versa.

Answer 2 (1996): Yes, and there is an efficient algorithm to do so!

Gödel prize (2003)

GENERAL BOOSTING RECIPE

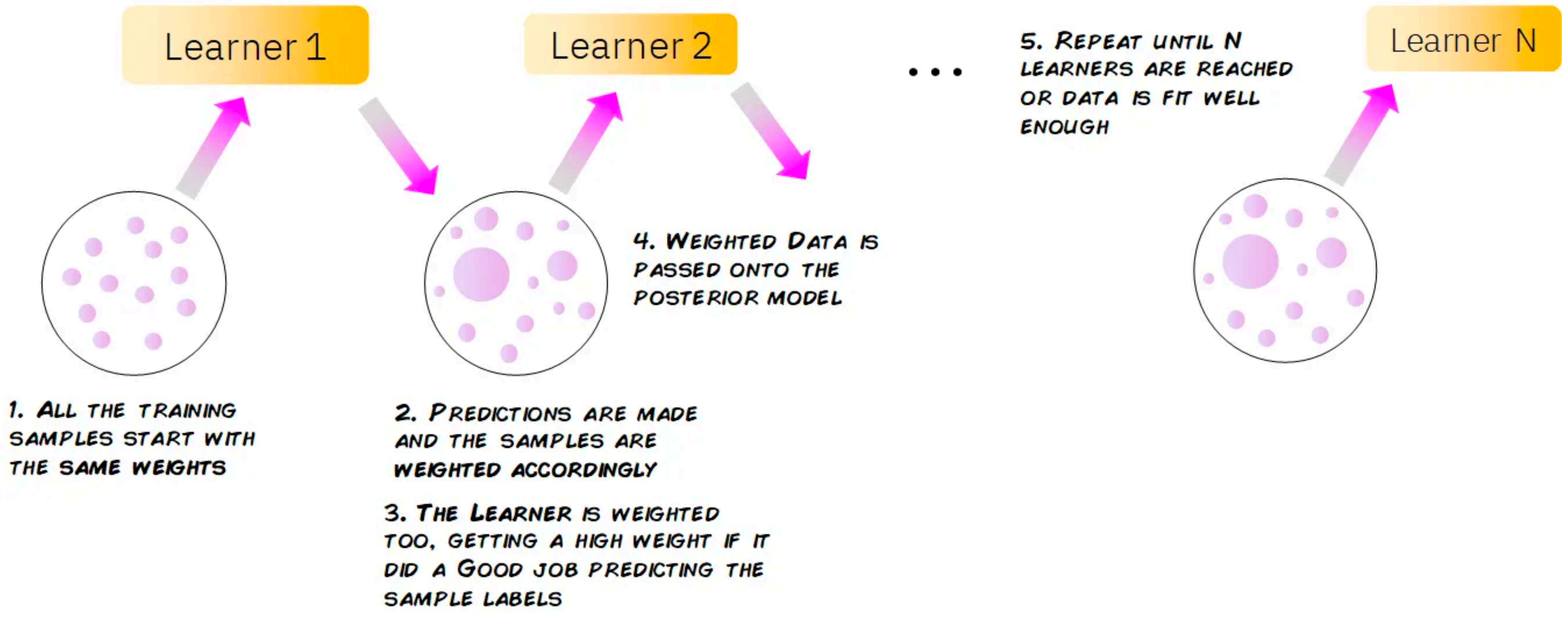


Image source: https://towardsdatascience.com/what-is-boosting-in-machine-learning-2244aa196682

BOOSTING VS BAGGING

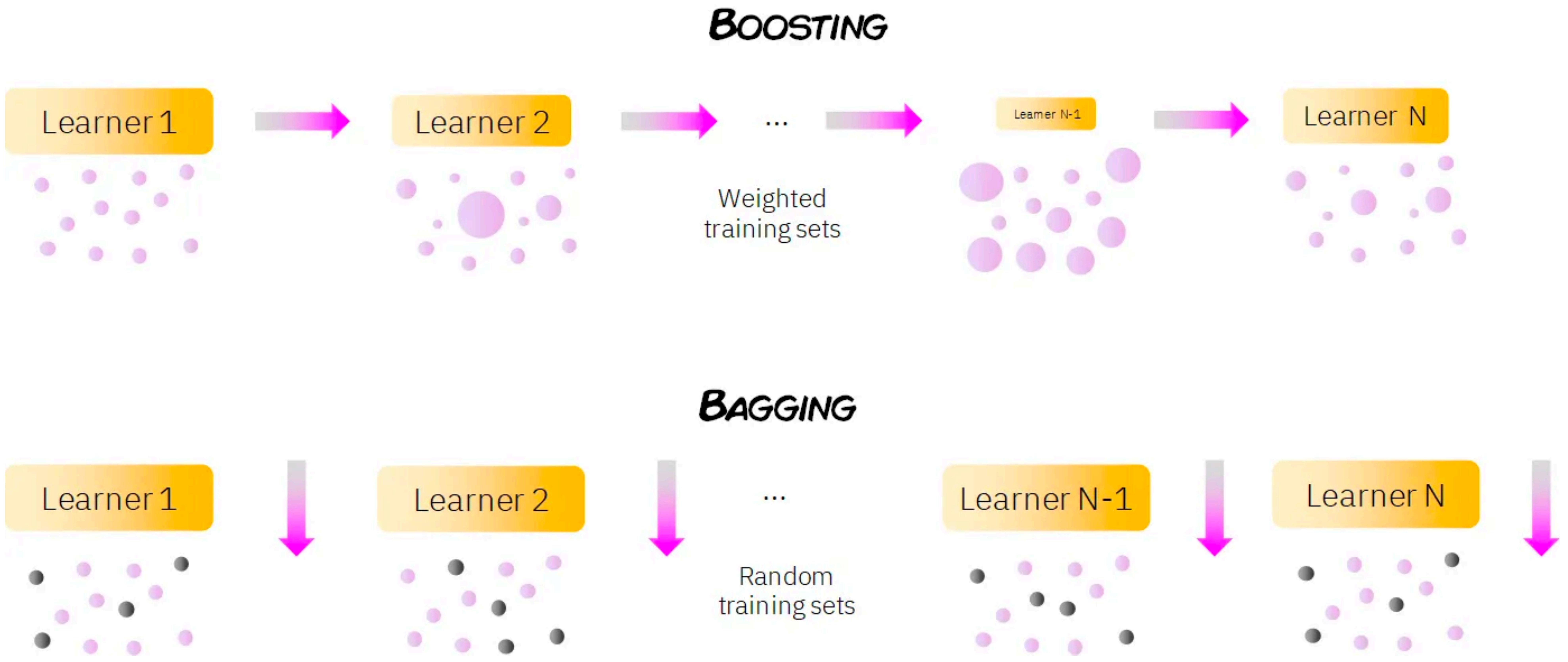


Image source: https://towardsdatascience.com/what-is-boosting-in-machine-learning-2244aa196682

GENERAL BOOSTING SCHEME

Weak learner $\mathscr{A}$ guarantees error $\leq 1/2 - \gamma$ for any distribution

Training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $y_i \in \{-1,1\}$

Algorithm 1: Generic Boosting Scheme

for $t = 1, 2, \ldots, T$ **do**
  Construct discrete distribution μ_t over $[m]$
  Run weak learner $\mathcal{A}$ on training data sampled according to μ_t to get classifier f_t with small error over μ_t, $\epsilon_t = \Pr_{i \sim \mu_t}[f_t(x_i) \neq y_i] = 1/2 - \gamma_t \leq 1/2 - \gamma$ (by weak learning assumption).
end
Output final classifier f constructed using $f_1, \ldots, f_T$.

Question 1: How do we choose μ_t?

Question 2: How do we construct final classifier f using $f_1, \ldots, f_T$?

ADABOOST - ADAPTIVE BOOSTING

Question 1: How do we choose μ_t?

For all $i \in [m]$,

$$\mu_{1,i} = \frac{1}{m}$$

Equal weight initially

$$\mu_{t+1,i} = \frac{\mu_{t,i}}{Z_t} \times \exp(-\alpha_t y_i f_t(x_i))$$

Normalizing factor

Weight increased if incorrect and decreased if correct

Optimal choice of shrinkage $\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ where $\epsilon_t = \Pr_{i\sim\mu_t}[f_t(x_i) \neq y_i]$.

ADABOOST - ADAPTIVE BOOSTING

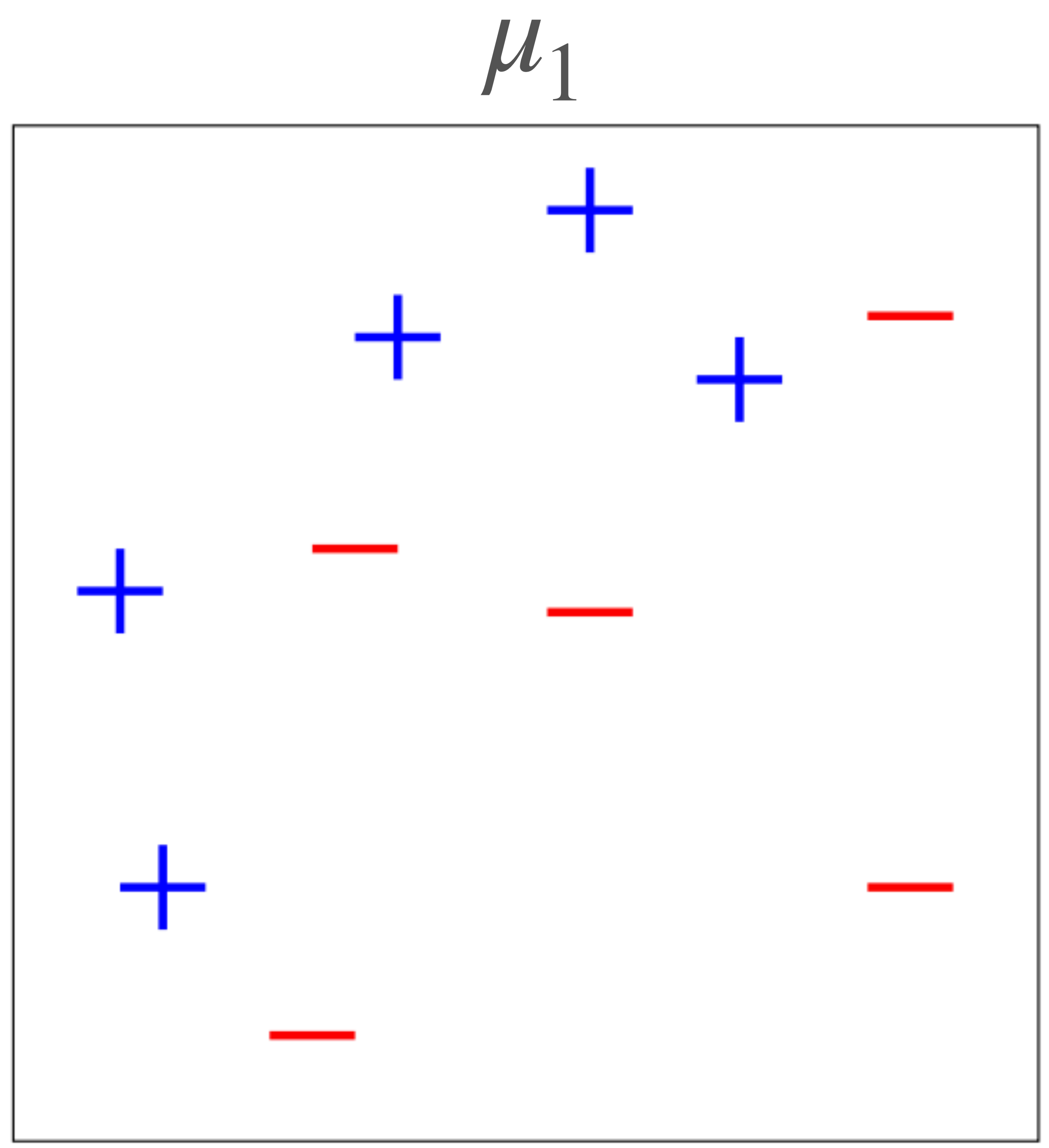
Question 2: How do we construct final classifier f using $f_1, \ldots, f_T$?

$$f(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t f_t(x)\right)$$

Weighted combination of the weak learners

The weight is based on how good the weak learner is

EXAMPLE BY FREUND AND SCHAPIRE



Weak Leaner - horizontal or vertical linear classifiers

Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

EXAMPLE BY FREUND AND SCHAPIRE

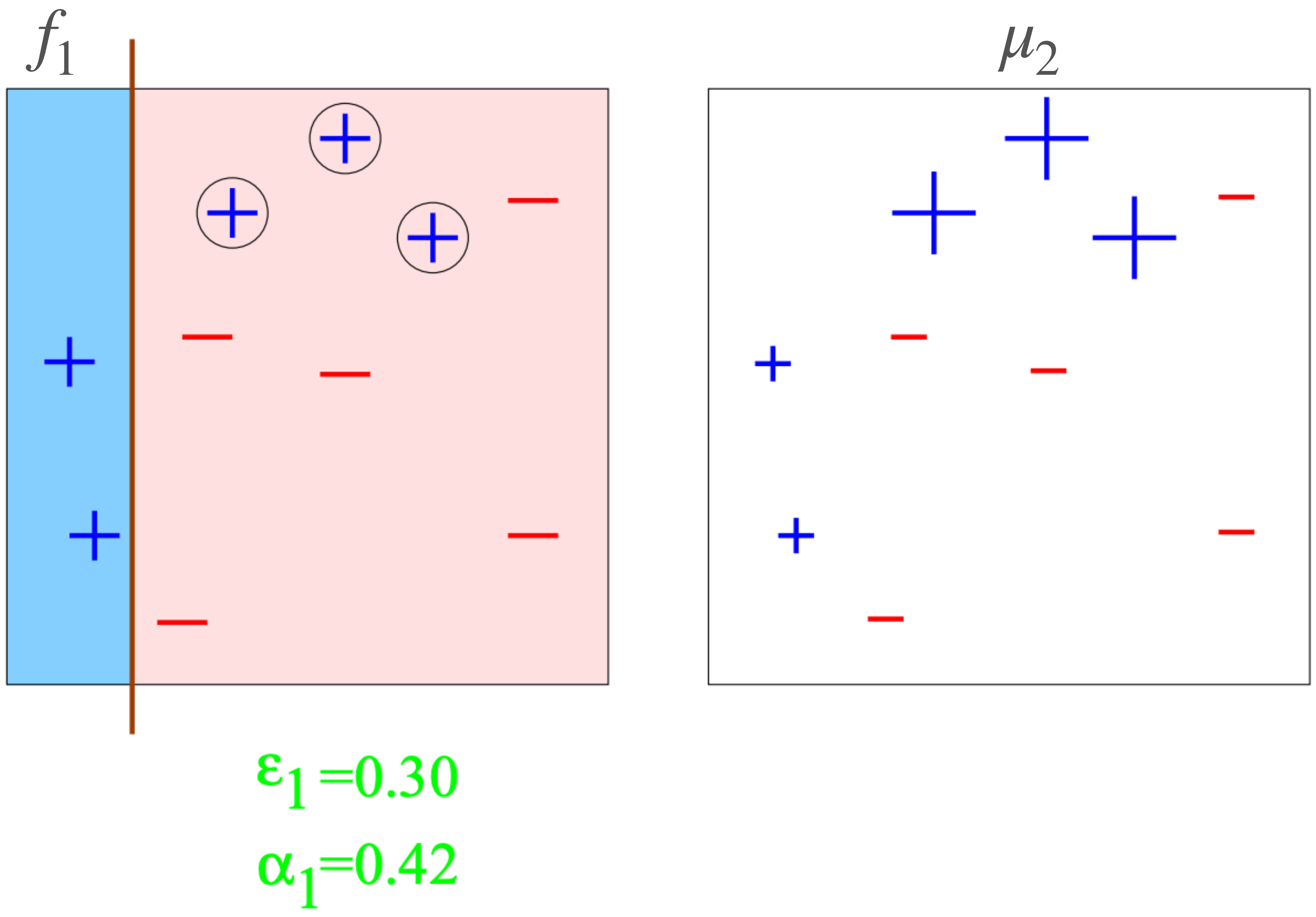


Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

EXAMPLE BY FREUND AND SCHAPIRE

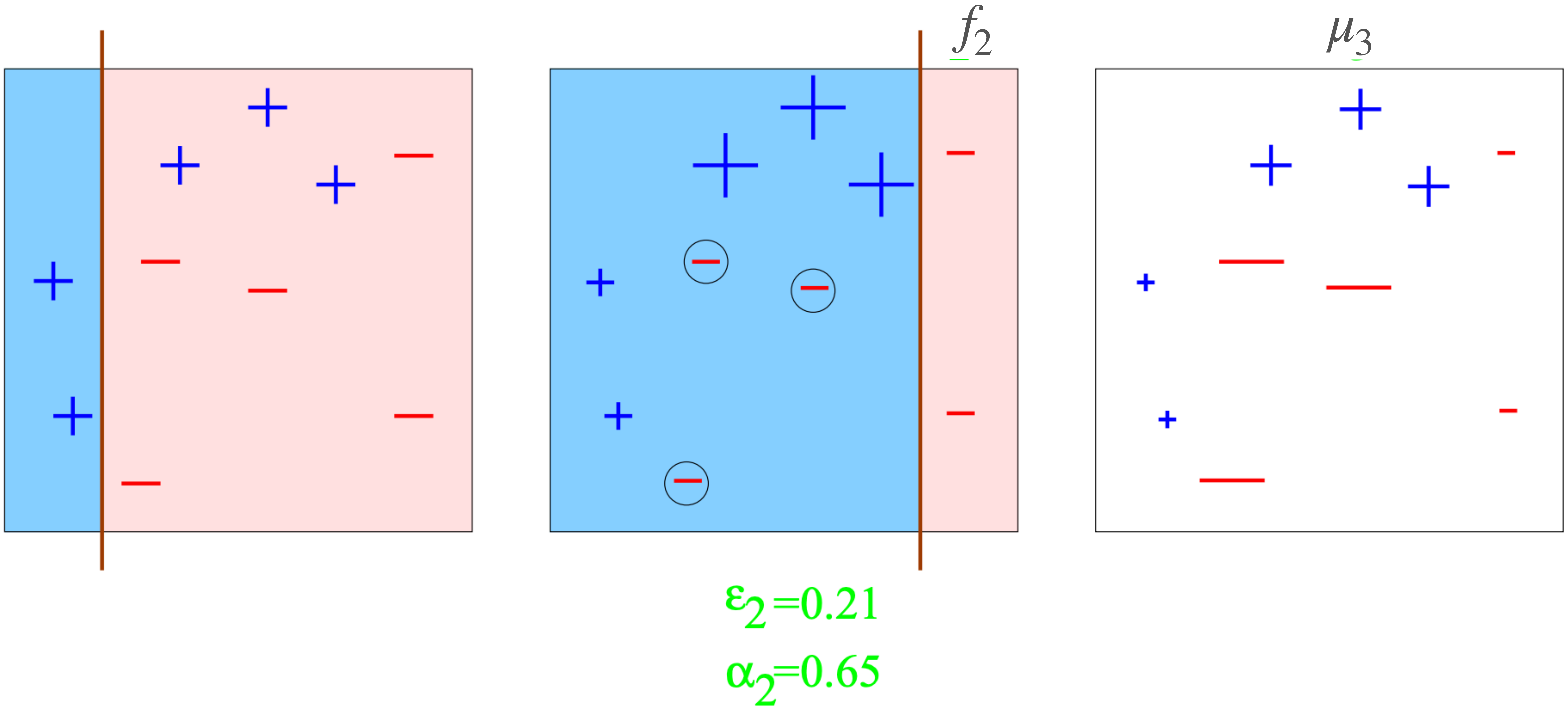


Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

EXAMPLE BY FREUND AND SCHAPIRE

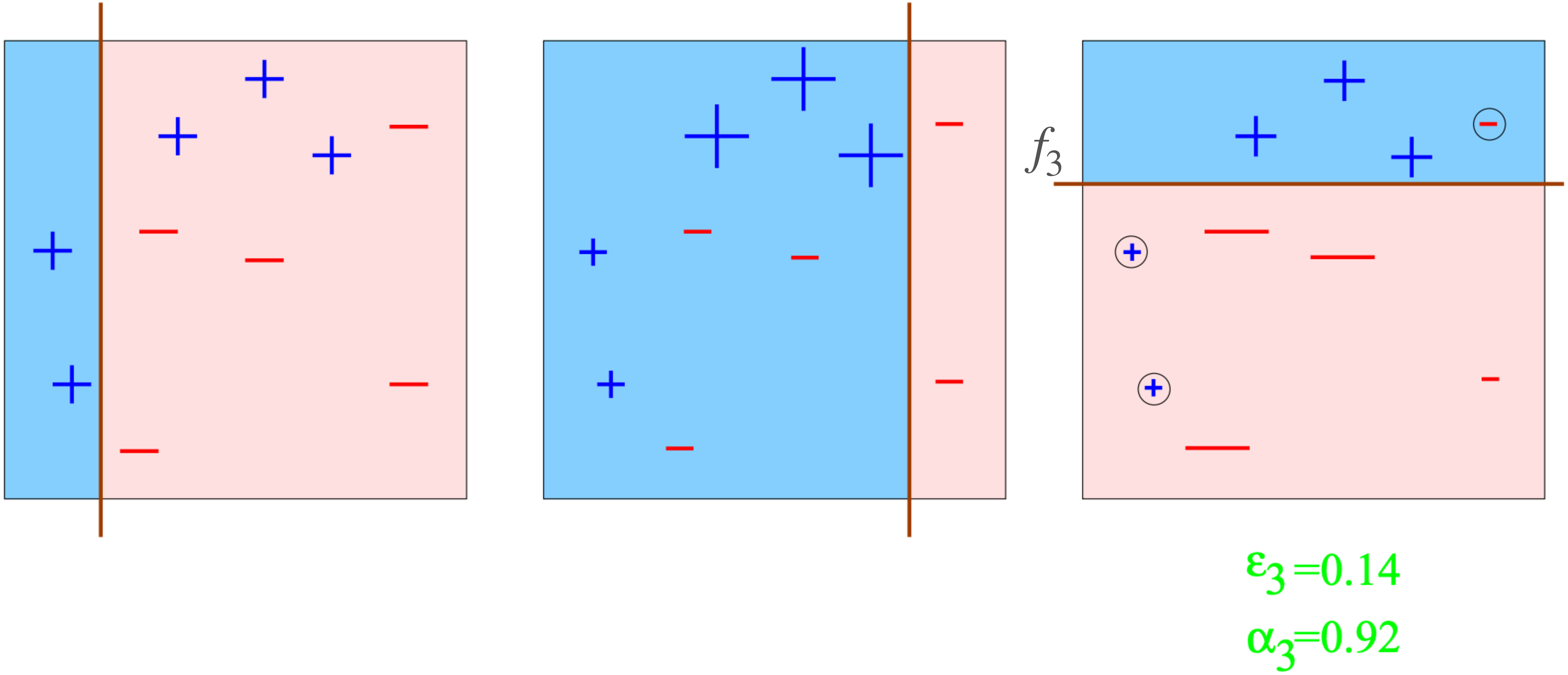


Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

EXAMPLE BY FREUND AND SCHAPIRE

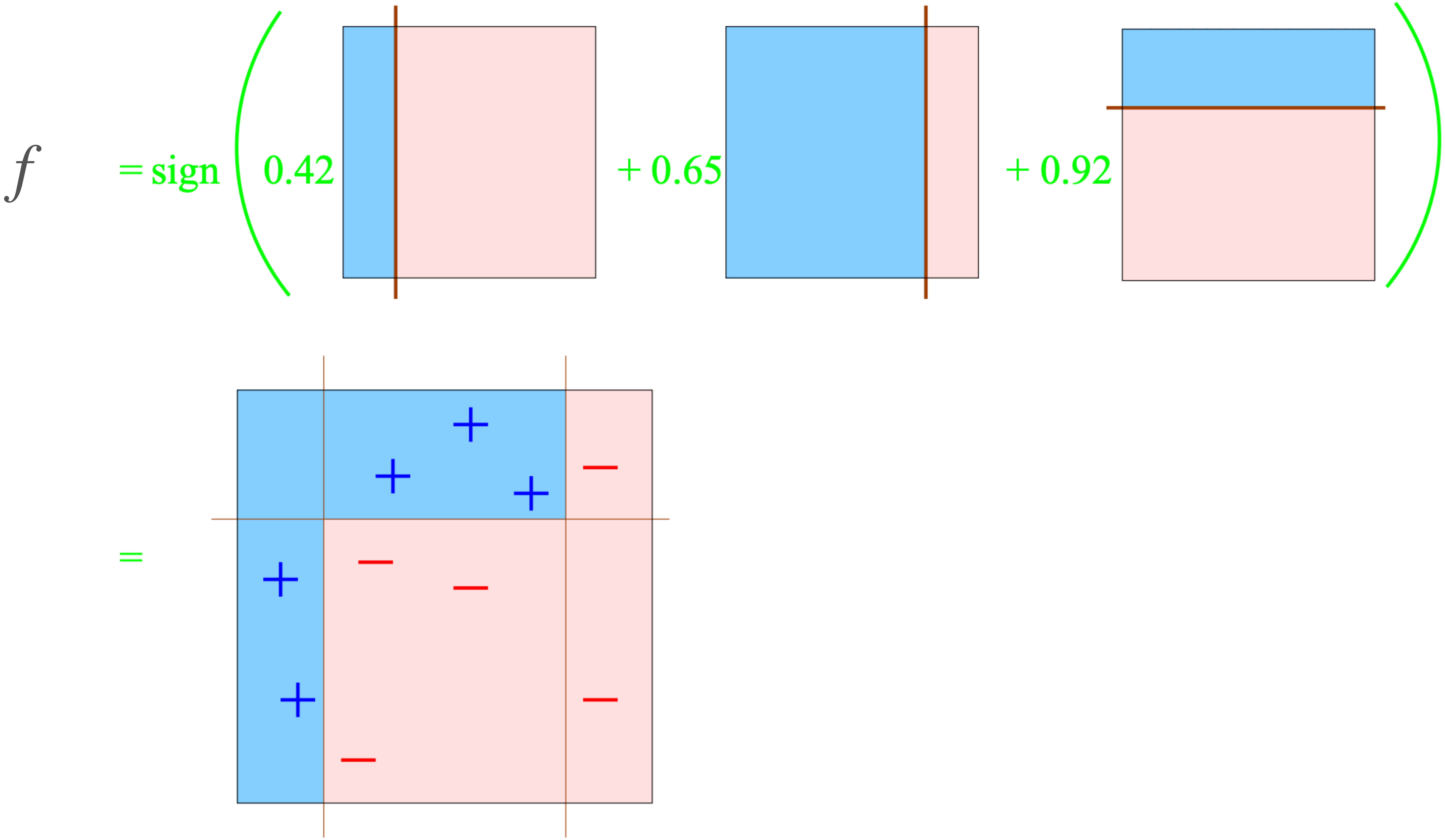


Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

TRAINING ERROR GUARANTEE

Weak learner $\mathscr{A}$ guarantees error $\leq 1/2 - \gamma$ for any distribution

Theorem:

Let f be the output of AdaBoost after T steps, then we have

$$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} 1[f(x_i) \neq y_i] \leq \exp\left(-2\gamma^2 T\right).$$

Training error goes down exponentially fast with the number of iterations

PROOF

Step 1: Bound on $\mu_{T+1,i}$

$$\mu_{T+1,i} = \frac{\exp\left(-y_i f(x_i)\right)}{m\prod_{t=1}^{T} Z_t}$$

Step 2: Bound on $\hat{R}(f)$

$$\hat{R}(f) \leq \prod_{t=1}^{T} Z_t$$

Step 3: Bound on Z_t

$$Z_t \leq \exp(-2\gamma^2)$$

Facts:

$$\mu_{1,i} = \frac{1}{m}$$

$$\mu_{t+1,i} = \frac{\mu_{t,i}}{Z_t} \times \exp(-\alpha_t y_i f_t(x_i))$$

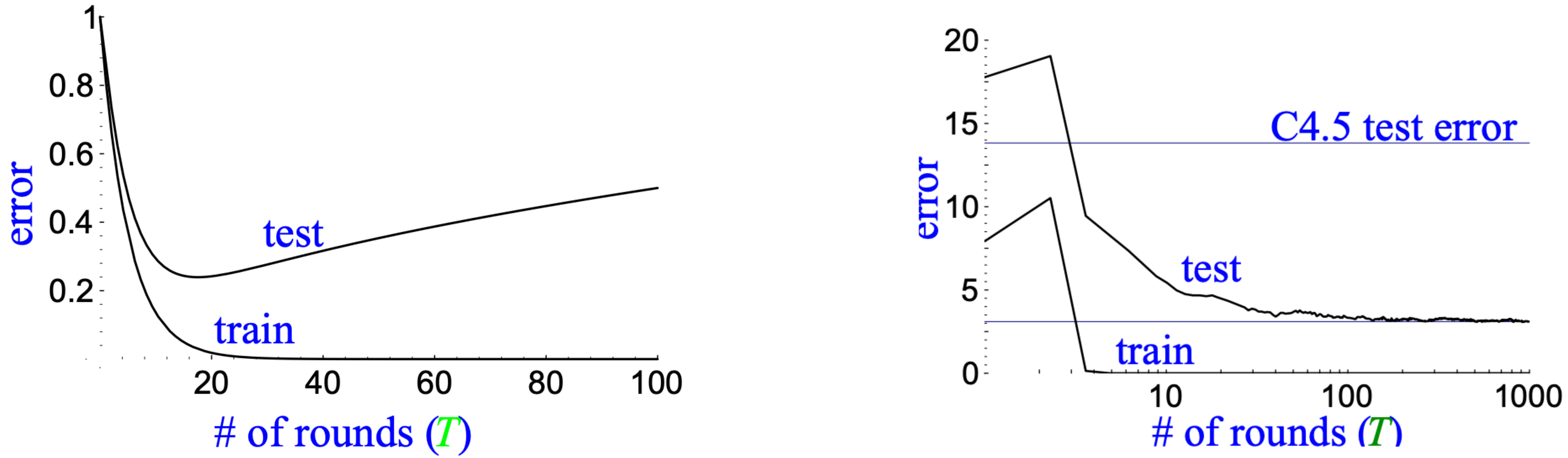
$$\hat{R}(f) = \frac{1}{m}\sum_{i=1}^{m} 1[f(x_i) \neq y_i]$$

$$Z_t = \sum_{j=1}^{m} \mu_{t,j} \exp(-\alpha_t y_j f_t(x_j))$$

GENERALIZATION PERFORMANCE

We reduced bias by creating a more complex classifier

What about the variance of the final classifier for increasing *T*?



Test error improves even after training error is 0!

Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

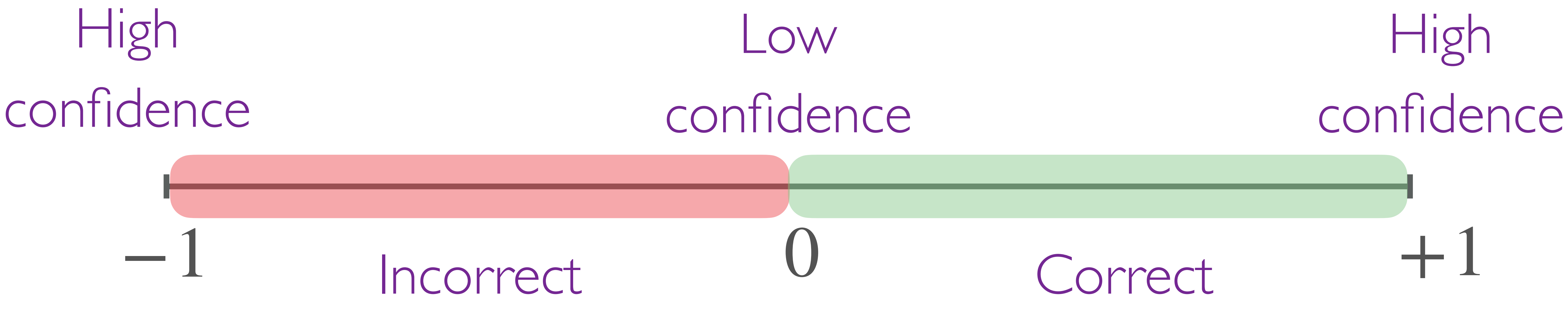
BIAS/VARIANCE - WHY NO TRADEOFF?

AdaBoost ensures large margin! (By Schapire, Freund, Bartlett & Lee)

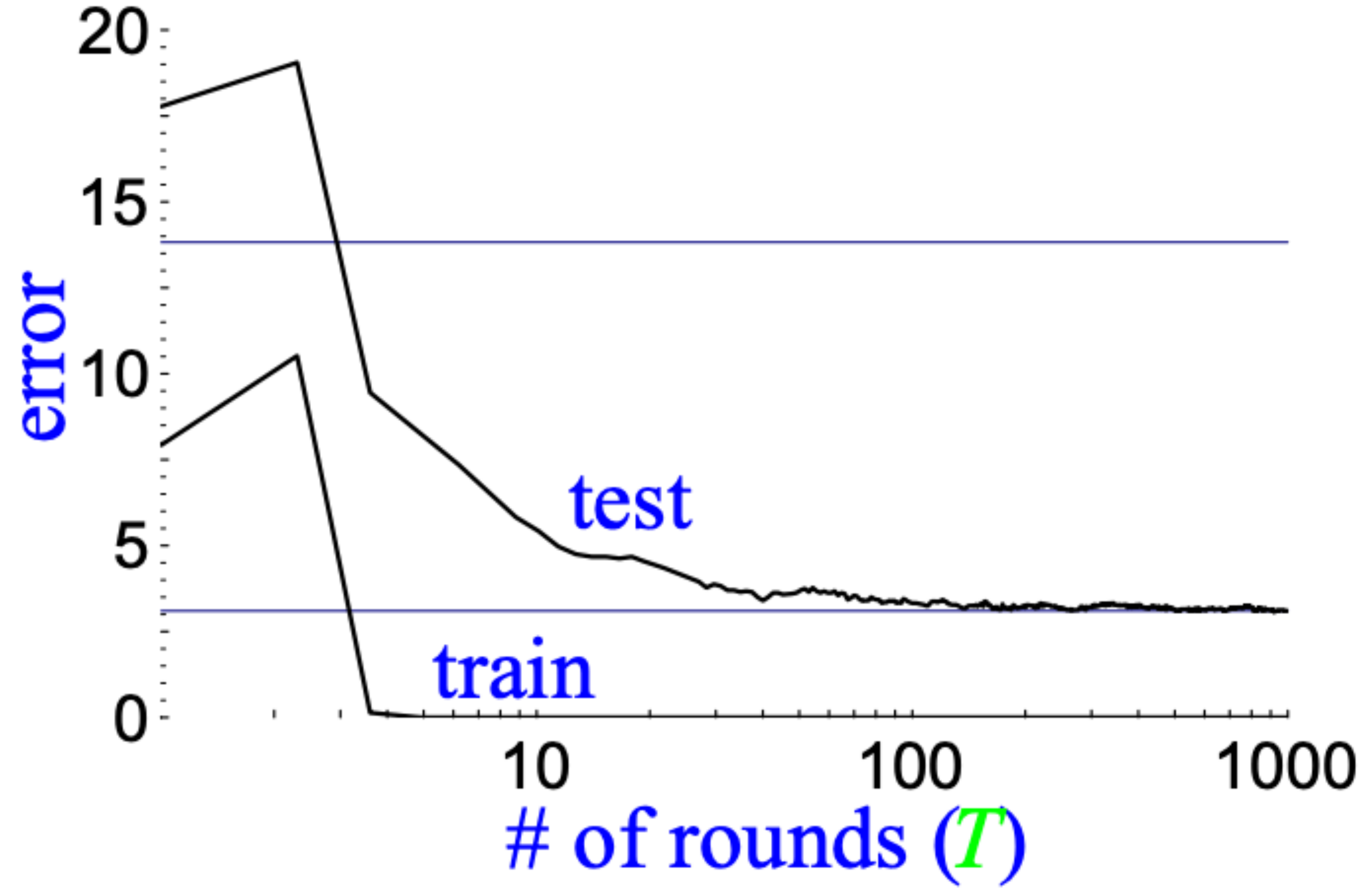
- Training error measures only correctness of prediction
- A better notion is confidence, how sure is the learner about the prediction
- In AdaBoost, the final classifier is a weighted vote of the weak learners

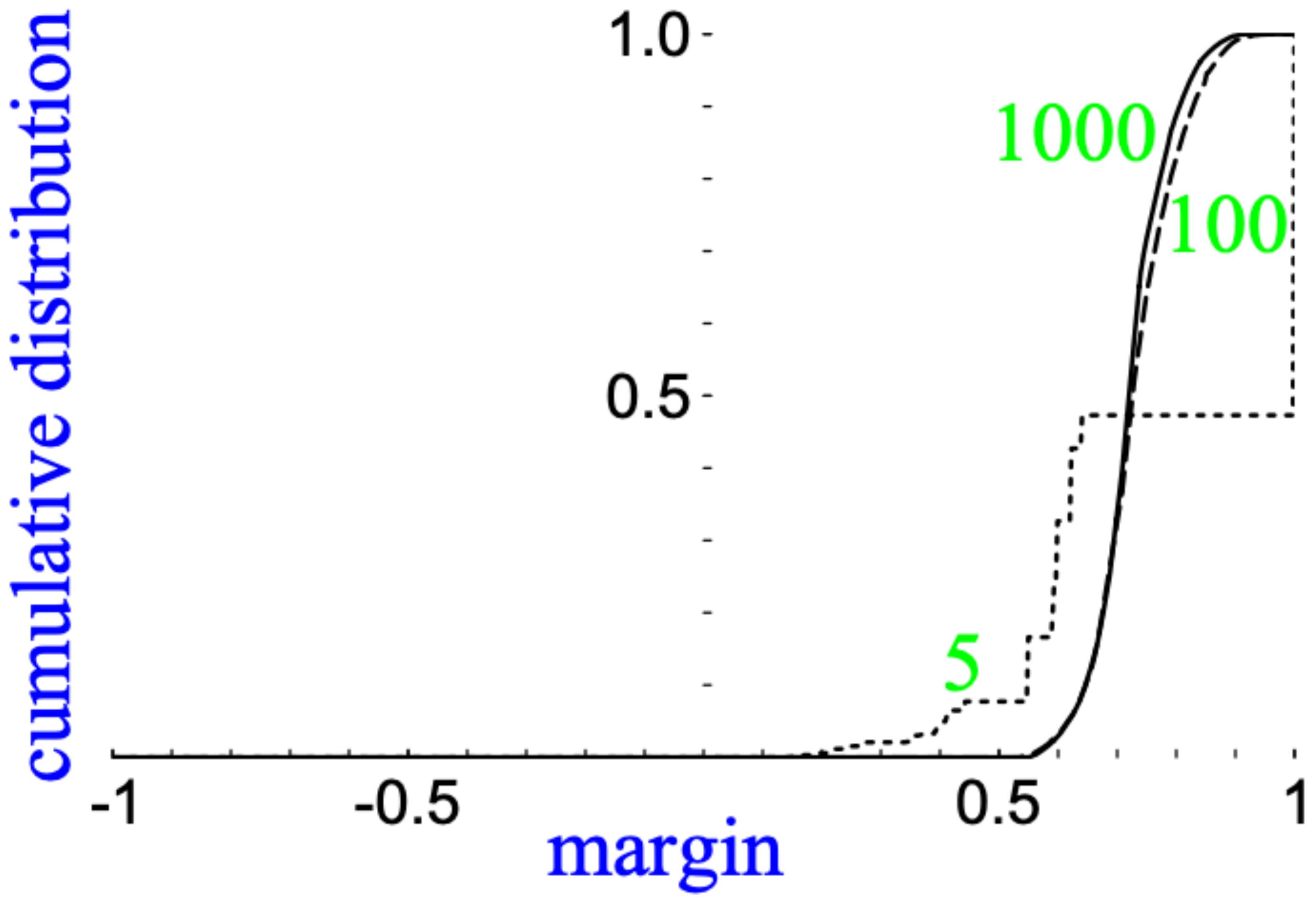
Margin - how strong is the vote?

= total weight of correct weak learners - total weight of incorrect weak learners



MARGIN





All points have margin at least 0.5

	# rounds		
	5	100	1000
train error	0.0	0.0	0.0
test error	8.4	3.3	3.1
% margins ≤ 0.5	7.7	0.0	0.0
minimum margin	0.14	0.52	0.55

Large margin $\Longrightarrow$ simpler classifier and better generalization

Image source: https://media.nips.cc/Conferences/2007/Tutorials/Slides/schapire-NIPS-07-tutorial.pdf

OPTIMIZATION VIEWPOINT OF BOOSTING

AdaBoost can be viewed as coordinate descent on a loss function over the space of linear combinations of weak classifiers

Recall that $\hat{R}(f) \leq \prod_{t=1}^{T} Z_t = \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i))$ where $f(x) = \sum_t \alpha_t f_t(x)$

- Coordinate descent would choose a coordinate and find the corresponding α to maximally decrease the loss
- AdaBoost is essentially doing coordinate descent on this loss
- An alternate way to view AdaBoost is via functional gradient descent

PROS AND CONS

Benefits of AdaBoost

- Fast
- Simple
- Only hyper-parameter is T
- Flexible - can use any weak learning algorithm
- Do not need to know how good the weak learner is
- Powerful - only weak learners needed

Caveats of AdaBoost

- Performance dependent on data and weak learner
- Can overfit if weak learner is too complex
- Can also underfit if weak learner is not good
- Not robust to noise