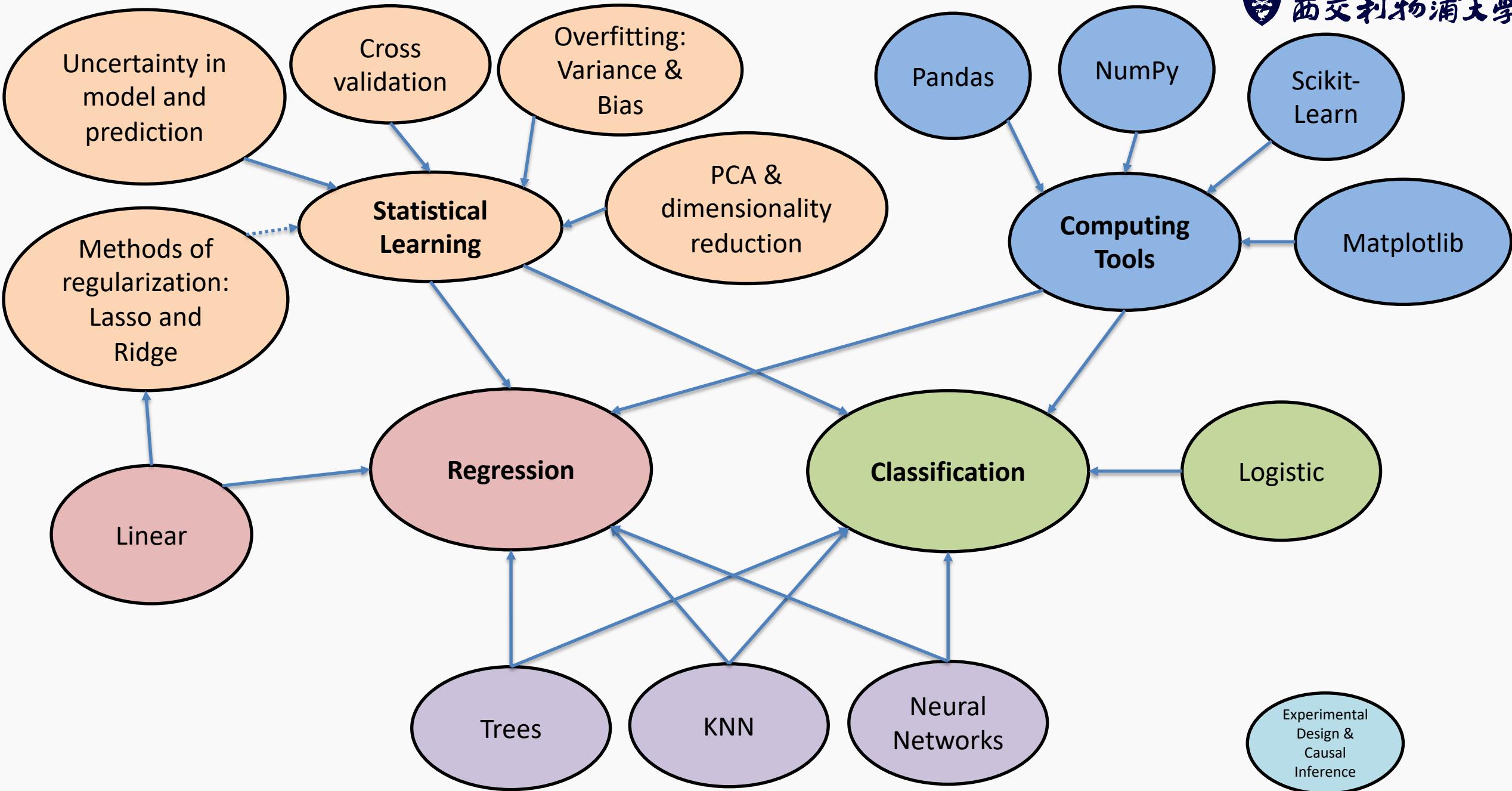


# INT 303 BIG DATA ANALYTICS

# Lecture12: Boosting

Pengfei FAN  
pengfei.fan@xjtu.edu.cn



# Outline

- Review of Ensemble Methods
  - Bags and Forests of Trees
- Boosting
  - Gradient Boosting
    - Set-up and intuition
    - Connection to Gradient Descent
    - The Algorithm
  - AdaBoost

# Bags and Forests of Trees

- Last time we examined how the short-comings of single decision tree models can be overcome by ensemble methods - making one model out of many trees.
- We focused on the problem of training large trees, these models have low bias but high variance.
- We compensated by training an ensemble of full decision trees and then averaging their predictions - thereby reducing the variance of our final model.

# Bags and Forests of Trees (cont.)

## Bagging:

- create an ensemble of trees, each trained on a bootstrap sample of the training set
- average the predictions.

## Random forest:

- create an ensemble of trees, each trained on a bootstrap sample of the training set
- in each tree and each split, randomly select a subset of predictors, choose a predictor from this subset for splitting
- average the predictions

Note that the ensemble building aspects of both methods are embarrassingly parallel!

# More

- Unbalance dataset
- Weighted samples
- Categorical data
- Missing data
- Why did we reject misclassification error?

# Boosting

# Motivation for Boosting

**Question:** Could we address the shortcomings of single decision trees models in some other way?

For example, rather than performing variance reduction on complex trees, can we decrease the bias of simple trees - make them more expressive?

Can we learn from our mistakes?

A solution to this problem, making an expressive model from simple trees, is another class of ensemble methods called ***boosting***.

# Boosting Algorithms

# Gradient Boosting

The key intuition behind boosting is that one can take an ensemble of simple models  $\{T_h\}_{h \in H}$  and additively combine them into a single, more complex model.

Each model  $T_h$  might be a poor fit for the data, but a linear combination of the ensemble

$$T = \sum_h \lambda_h T_h$$

can be expressive/flexible.

**Question:** But which models should we include in our ensemble? What should the coefficients or weights in the linear combination be?

# Gradient Boosting: the algorithm

**Gradient boosting** is a method for iteratively building a complex regression model  $T$  by adding simple models. Each new simple model added to the ensemble compensates for the weaknesses of the current ensemble.

1. Fit a simple model  $T^{(0)}$  on the training data

$$\{(x_1, y_1), \dots, (x_N, y_N)\}$$

Set  $T \leftarrow T^{(0)}$ . Compute the residuals  $\{r_1, \dots, r_N\}$  for  $T$ .

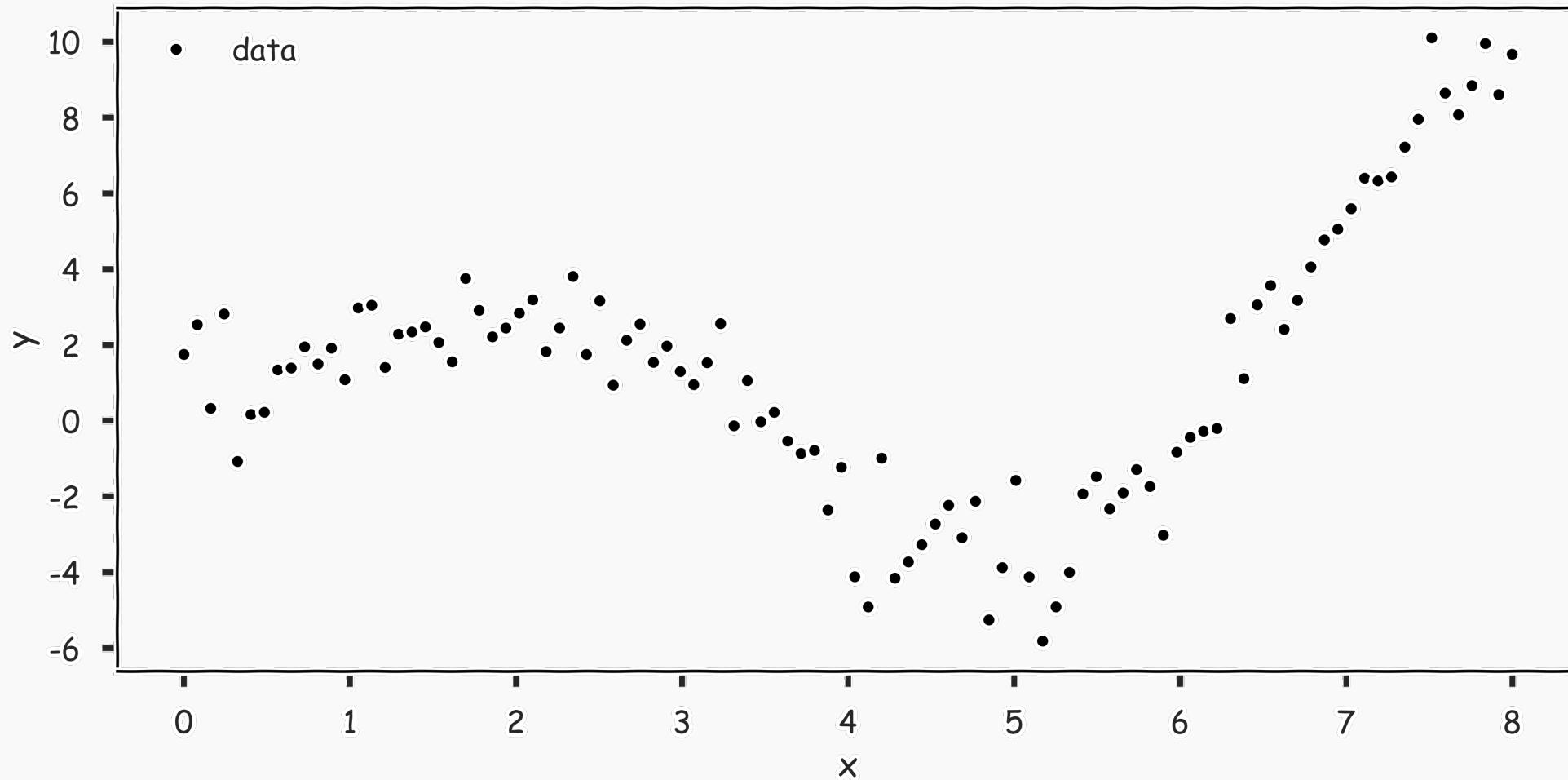
2. Fit a simple model,  $T^{(1)}$ , to the current **residuals**, i.e. train using

$$\{(x_1, r_1), \dots, (x_N, r_N)\}$$

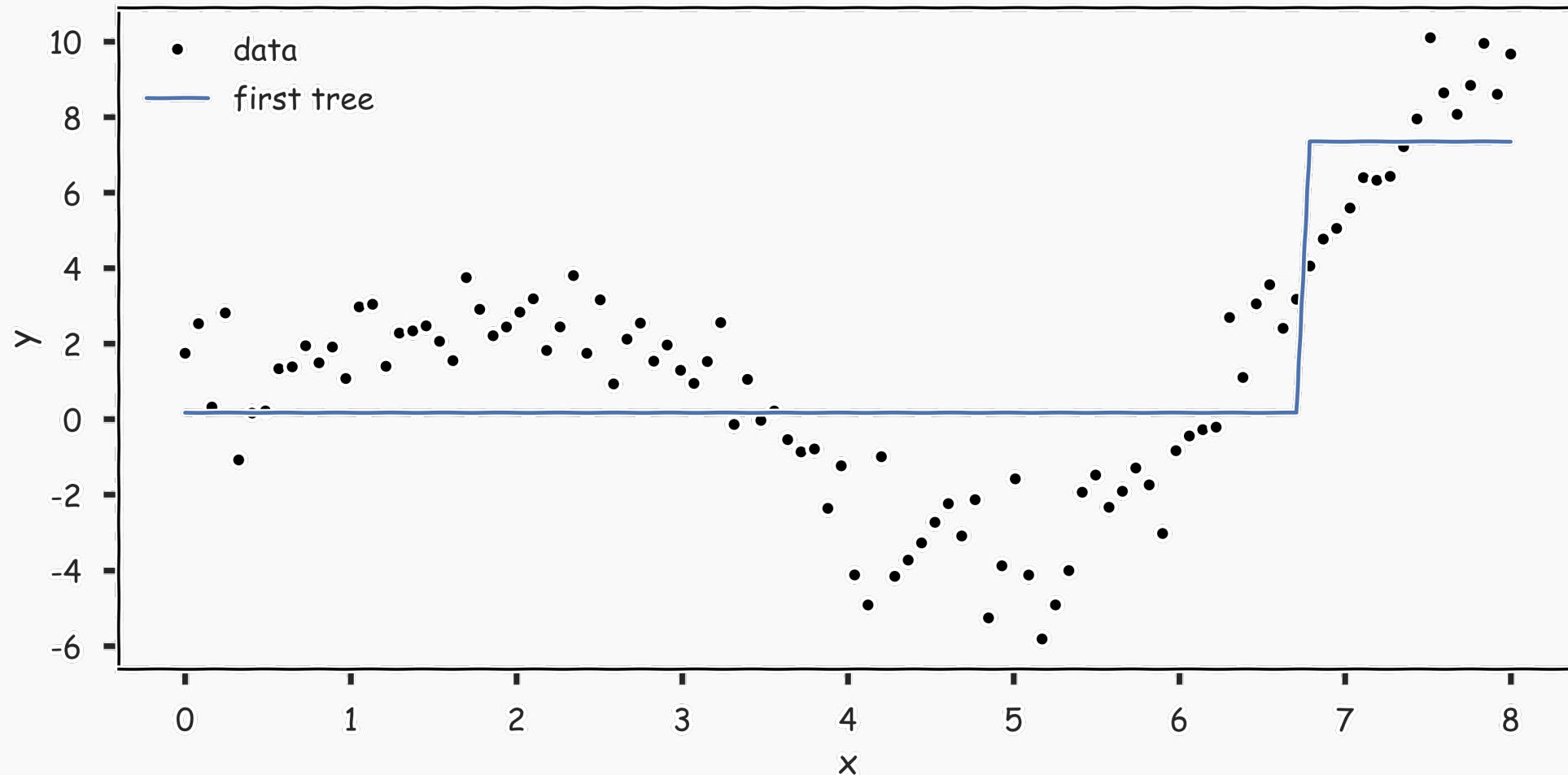
3. Set  $T \leftarrow T + \lambda T^{(1)}$
4. Compute residuals, set  $r_n \leftarrow r_n - \lambda T^i(x_n)$ ,  $n = 1, \dots, N$
5. Repeat steps 2-4 until **stopping** condition met.

where  $\lambda$  is a constant called the **learning rate**.

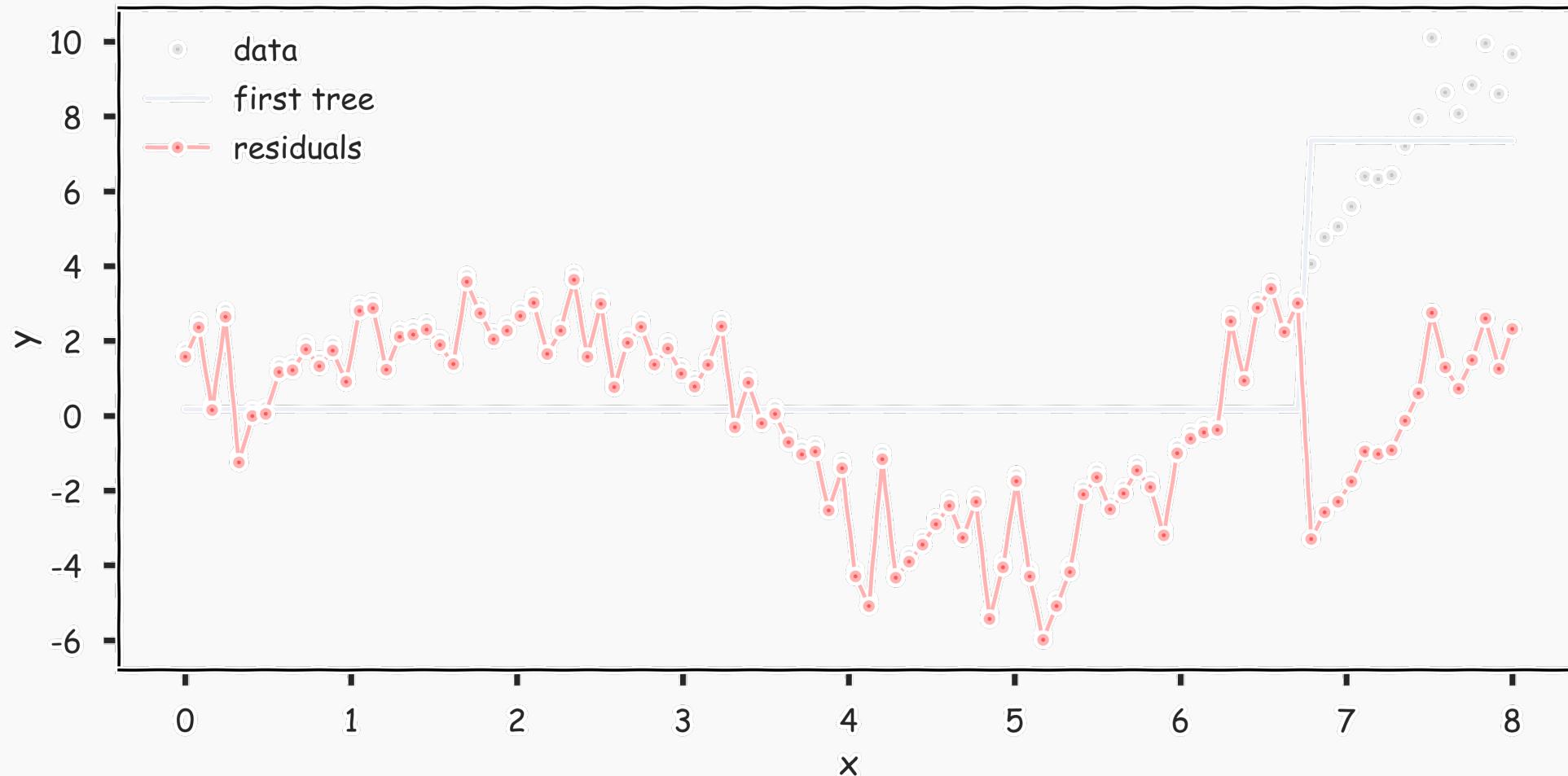
# Gradient Boosting: illustration



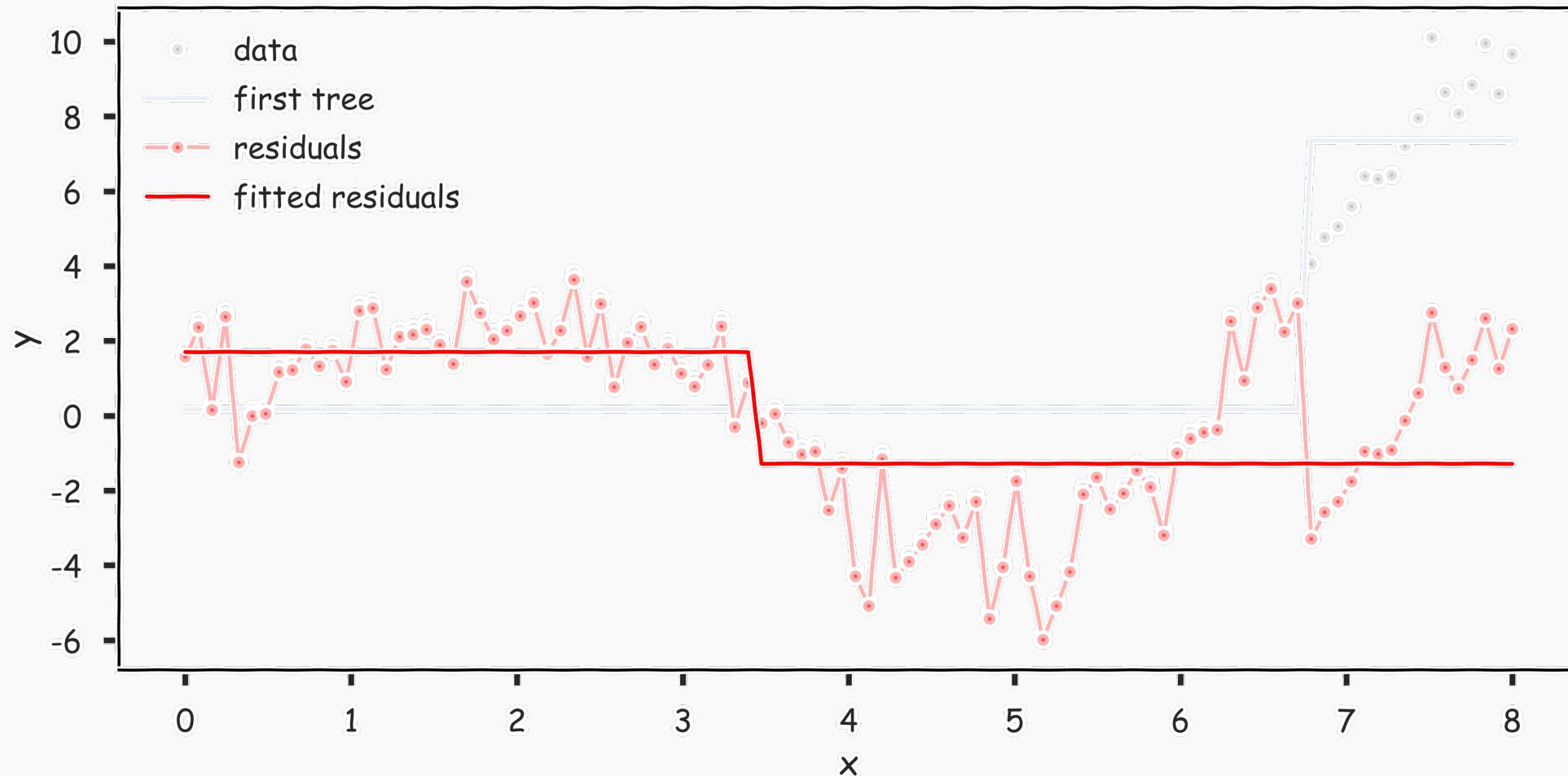
# Gradient Boosting: illustration



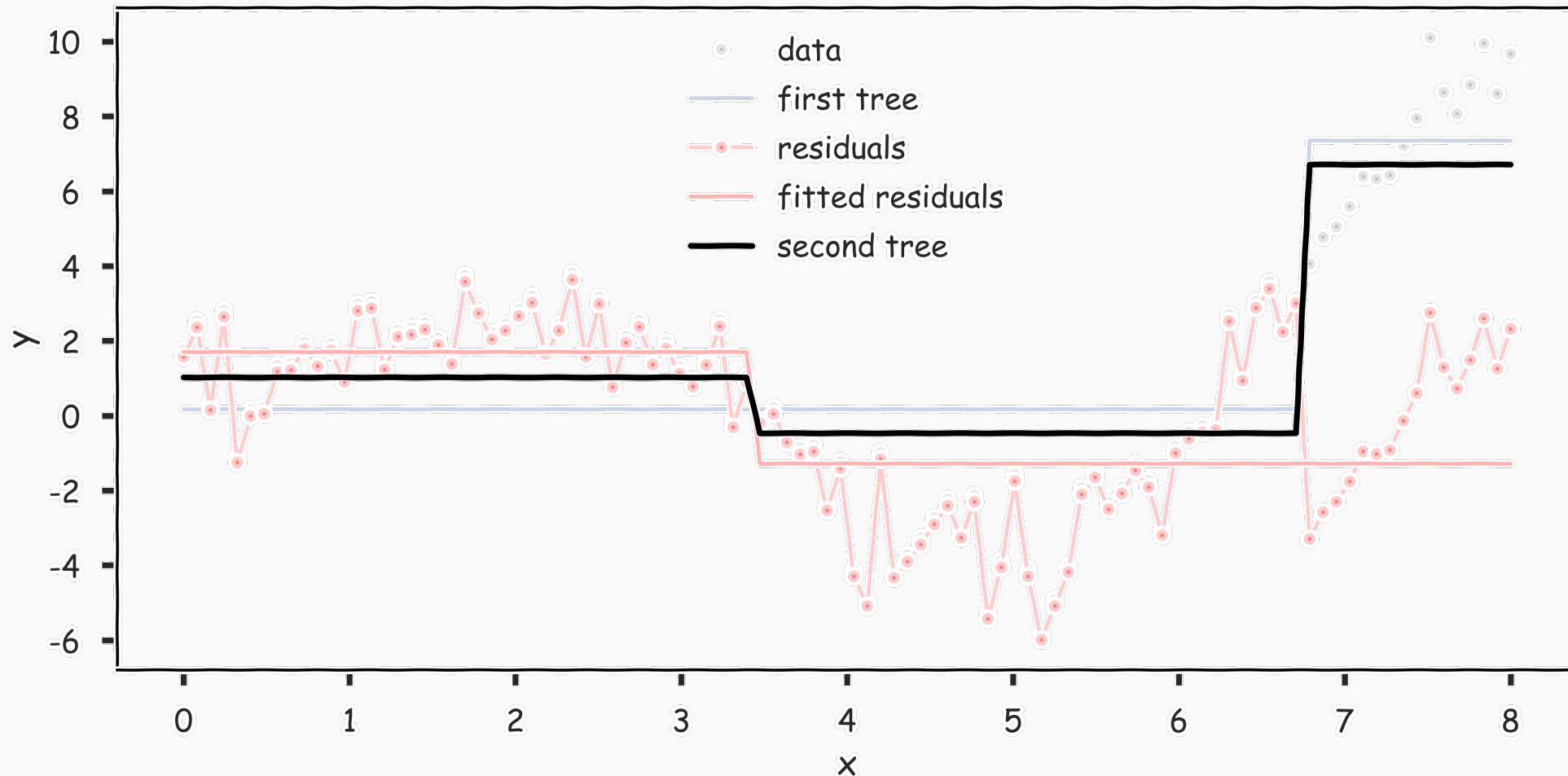
# Gradient Boosting: illustration



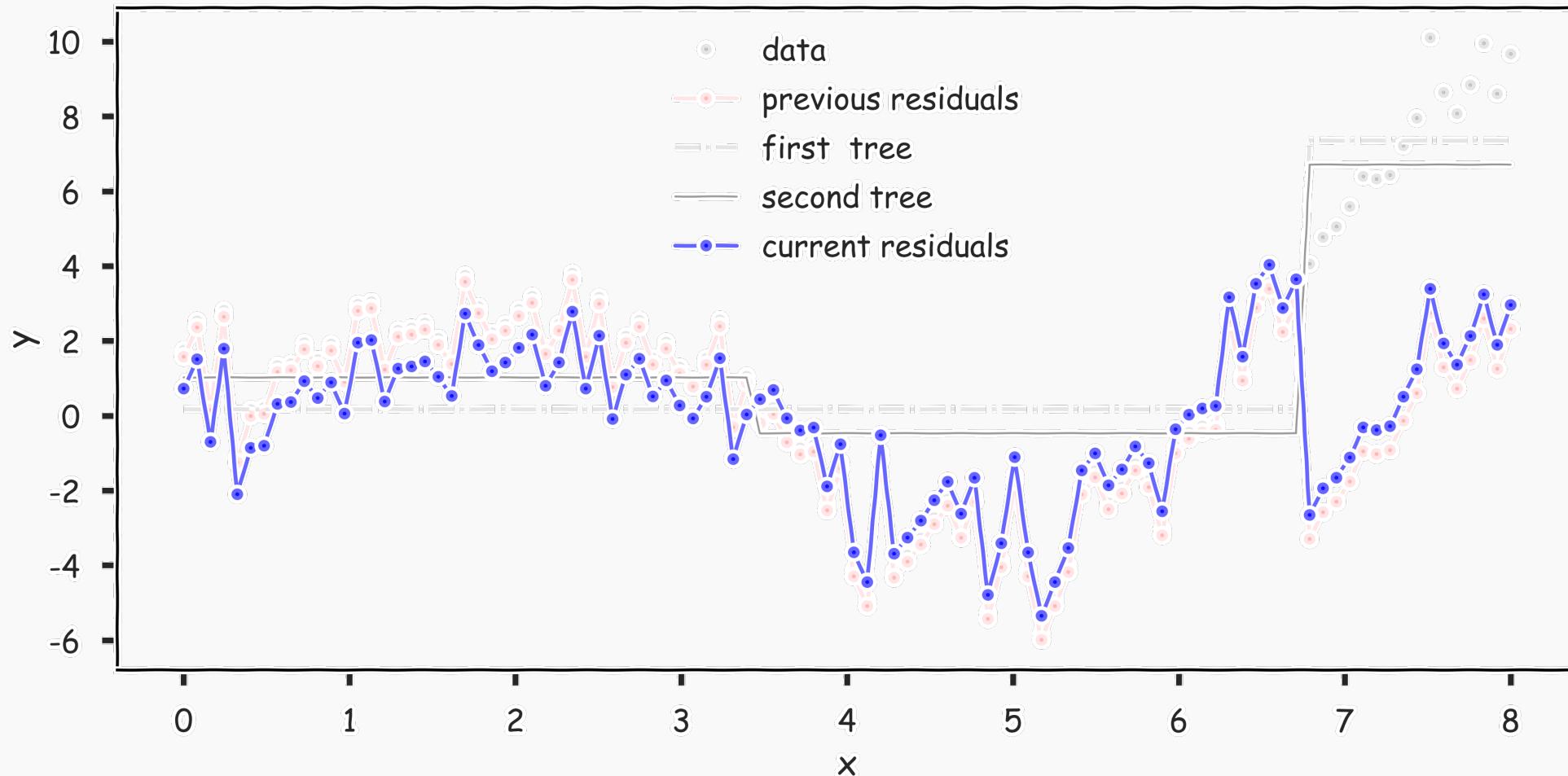
# Gradient Boosting: illustration



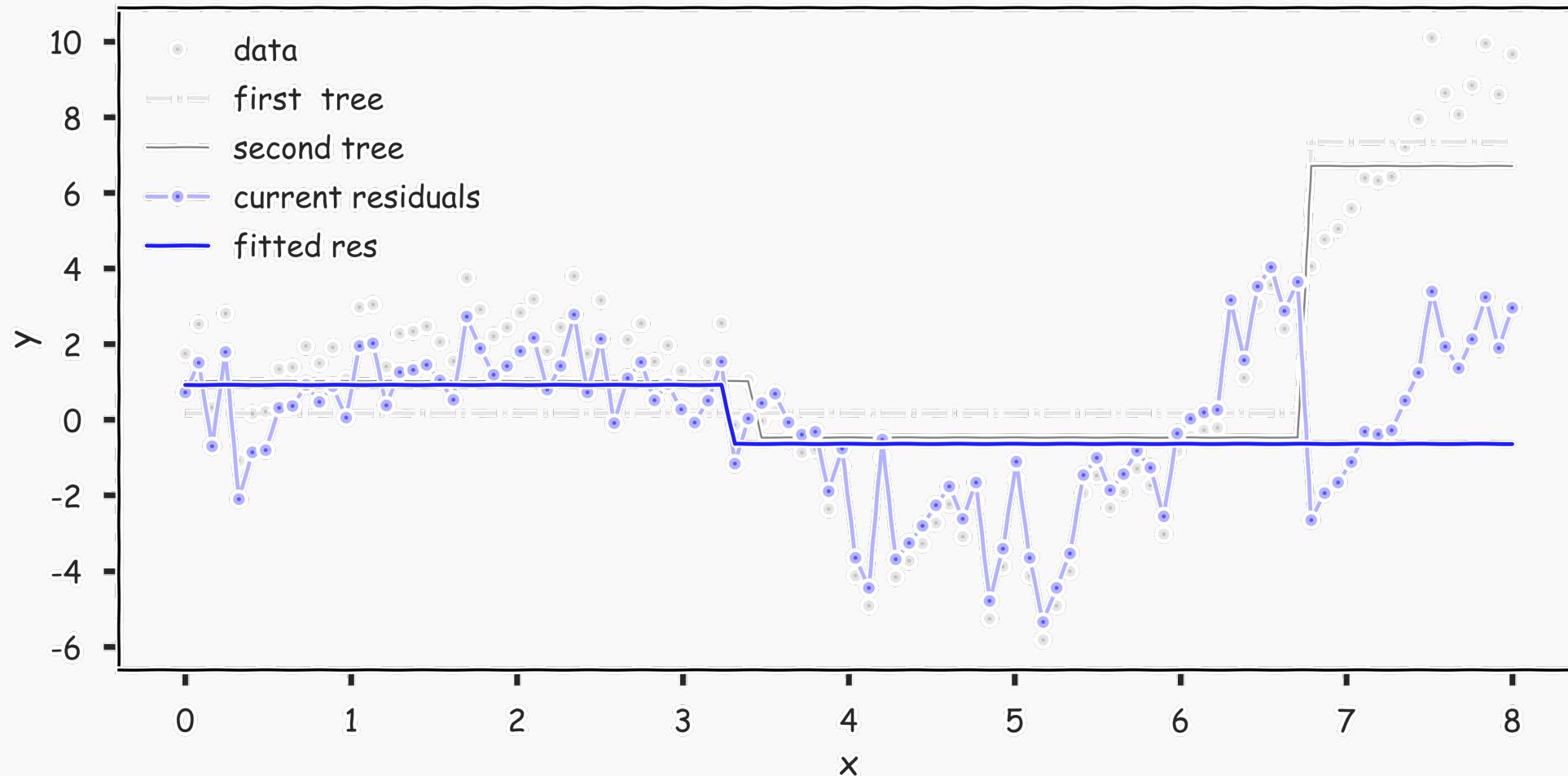
# Gradient Boosting: illustration



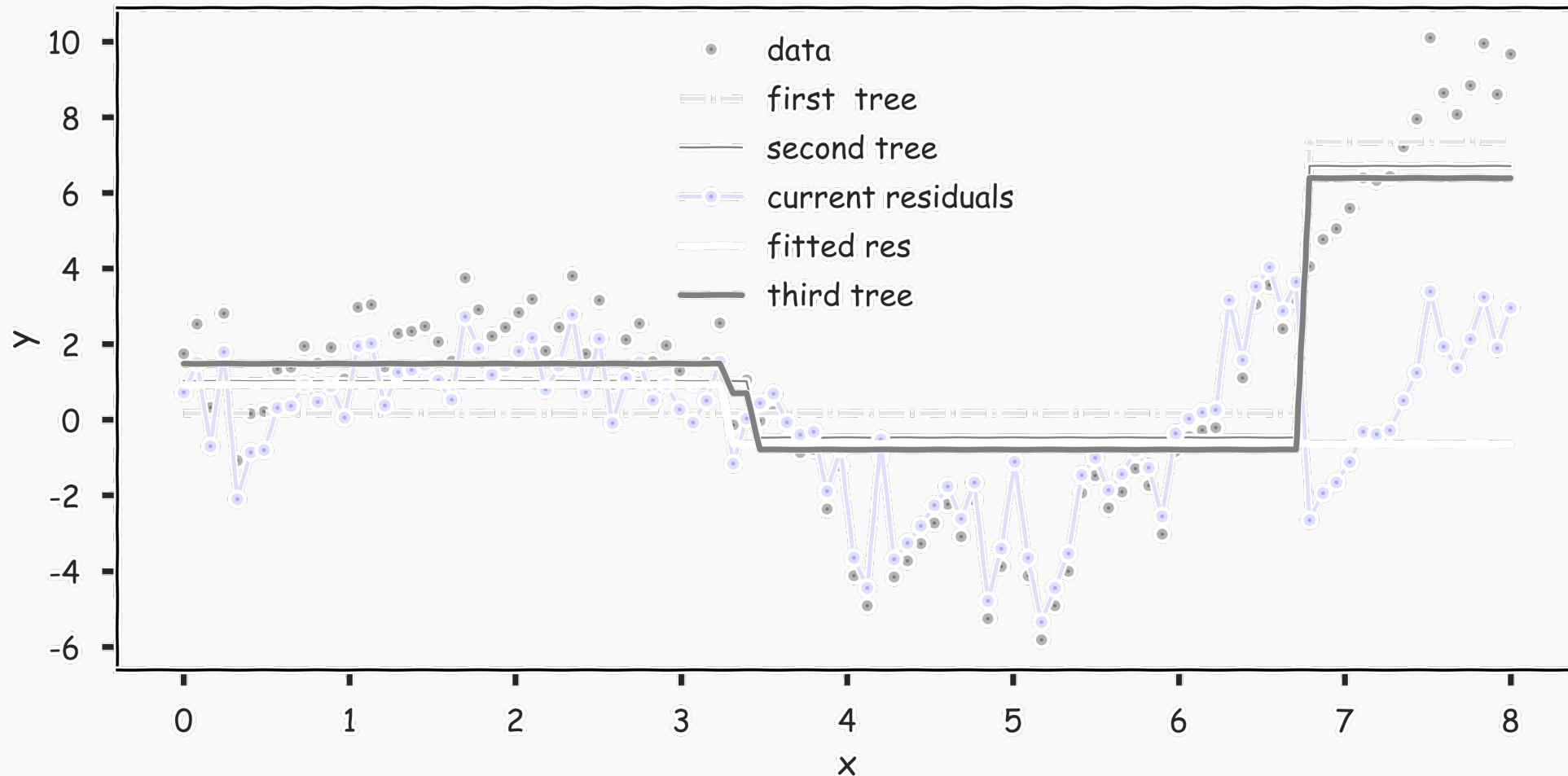
# Gradient Boosting: illustration



# Gradient Boosting: illustration



# Gradient Boosting: illustration



# Why Does Gradient Boosting Work?

Intuitively, each simple model  $T^{(i)}$  we add to our ensemble model  $T$ , models the errors of  $T$ .

Thus, with each addition of  $T^{(i)}$ , the residual is reduced

$$r_n - \lambda T^{(i)}(x_n)$$

**Note** that gradient boosting has a tuning parameter,  $\lambda$ .

If we want to easily reason about how to choose  $\lambda$  and investigate the effect of  $\lambda$  on the model  $T$ , we need a bit more mathematical formalism. In particular, how can we effectively descend through this optimization via an iterative algorithm?

We need to formulate gradient boosting as a type of *gradient descent*.

# Review: A Brief Sketch of Gradient Descent

In optimization, when we wish to minimize a function, called the *objective function*, over a set of variables, we compute the partial derivatives of this function with respect to the variables.

If the partial derivatives are sufficiently simple, one can analytically find a common root - i.e. a point at which all the partial derivatives vanish; this is called a *stationary point*.

If the objective function has the property of being *convex*, then the stationary point is precisely the min.

# Review: A Brief Sketch of Gradient Descent the Algorithm

In practice, our objective functions are complicated and analytically find the stationary point is intractable.

Instead, we use an iterative method called *gradient descent*:

1. Initialize the variables at any value:

$$x = [x_1, \dots, x_J]$$

2. Take the gradient of the objective function at the current variable values:

$$\nabla f(x) = \left[ \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_J}(x) \right]$$

3. Adjust the variables values by some negative multiple of the gradient:

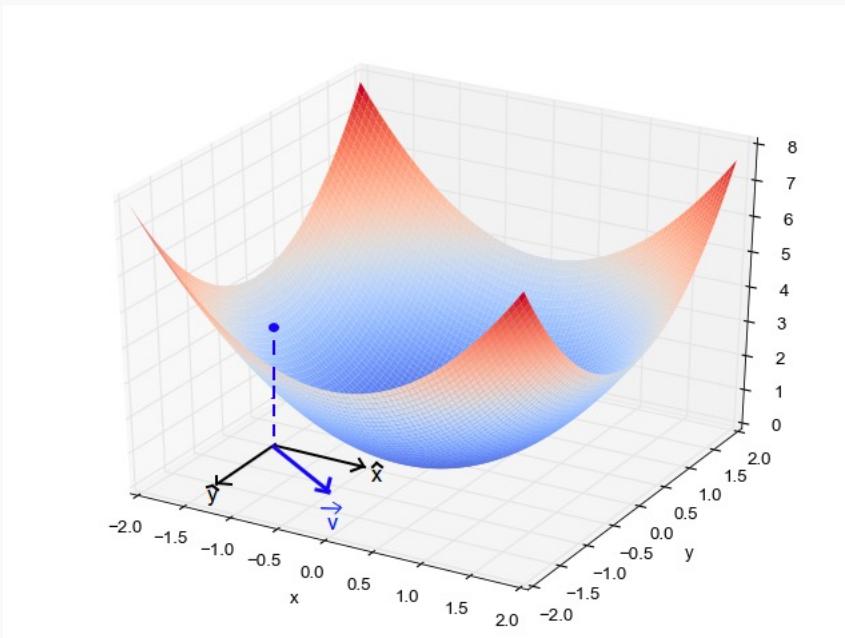
$$x \leftarrow x - \lambda \nabla f(x)$$

The factor  $\lambda$  is often called the learning rate.

# Why Does Gradient Descent Work?

**Claim:** If the function is convex, this iterative methods will eventually move  $x$  close enough to the minimum, for an appropriate choice of  $\lambda$ .

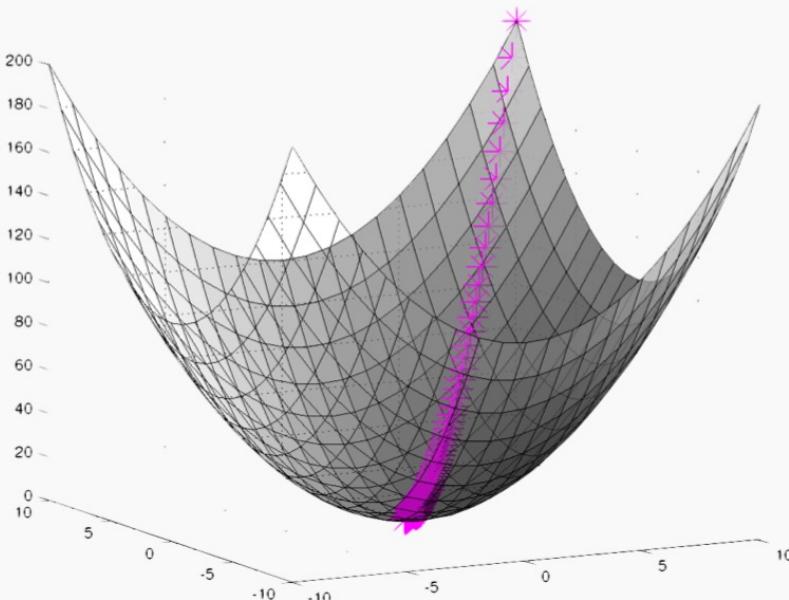
**Why does this work?** Recall, that as a vector, the gradient at point gives the direction for the greatest possible rate of increase.



# Why Does Gradient Descent Work?

Subtracting a  $\lambda$  multiple of the gradient from  $x$ , moves  $x$  in the *opposite* direction of the gradient (hence towards the steepest decline) by a step of size  $\lambda$ .

If  $f$  is convex, and we keep taking steps descending on the graph of  $f$ , we will eventually reach the minimum.



# Gradient Boosting as Gradient Descent

Often in regression, our objective is to minimize the MSE

$$\text{MSE}(\hat{y}_1, \dots, \hat{y}_N) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Treating this as an optimization problem, we can try to directly minimize the MSE with respect to the predictions

$$\begin{aligned}\nabla \text{MSE} &= \left[ \frac{\partial \text{MSE}}{\partial \hat{y}_1}, \dots, \frac{\partial \text{MSE}}{\partial \hat{y}_N} \right] \\ &= -2 [y_1 - \hat{y}_1, \dots, y_N - \hat{y}_N] \\ &= -2 [r_1, \dots, r_N]\end{aligned}$$

The update step for gradient descent would look like

$$\hat{y}_n \leftarrow \hat{y}_n + \lambda r_n, \quad n = 1, \dots, N$$

# Gradient Boosting as Gradient Descent (cont.)

There are two reasons why minimizing the MSE with respect to  $\hat{y}_n$ 's is not interesting:

- We know where the minimum MSE occurs:  $\hat{y}_n = y_n$ , for every  $n$ .
- Learning sequences of predictions,  $\hat{y}_n^1, \dots, \hat{y}_n^i, \dots$ , does not produce a model. The predictions in the sequences do not depend on the predictors!

# Gradient Boosting as Gradient Descent (cont.)

$$\hat{y}_n \leftarrow \hat{y}_n + \lambda r_n$$

The solution is to change the update step in gradient descent. Instead of using the gradient - the residuals - we use an *approximation* of the gradient that depends on the predictors:

$$\hat{y} \leftarrow \hat{y}_n + \lambda \hat{r}_n(x_n), \quad n = 1, \dots, N$$

In gradient boosting, we use a simple model to approximate the residuals,  $\hat{r}_n(x_n)$ , in each iteration.

**Motto:** gradient boosting is a form of gradient descent with the MSE as the objective function.

**Technical note:** note that gradient boosting is descending in a space of models or functions relating  $x_n$  to  $y_n$ !

# Gradient Boosting as Gradient Descent (cont.)

But why do we care that gradient boosting is gradient descent?

By making this connection, we can import the massive amount of techniques for studying gradient descent to analyze gradient boosting.

**For example**, we can easily reason about how to choose the learning rate  $\lambda$  in gradient boosting.

# Choosing a Learning Rate

Under ideal conditions, gradient descent iteratively approximates and converges to the optimum.

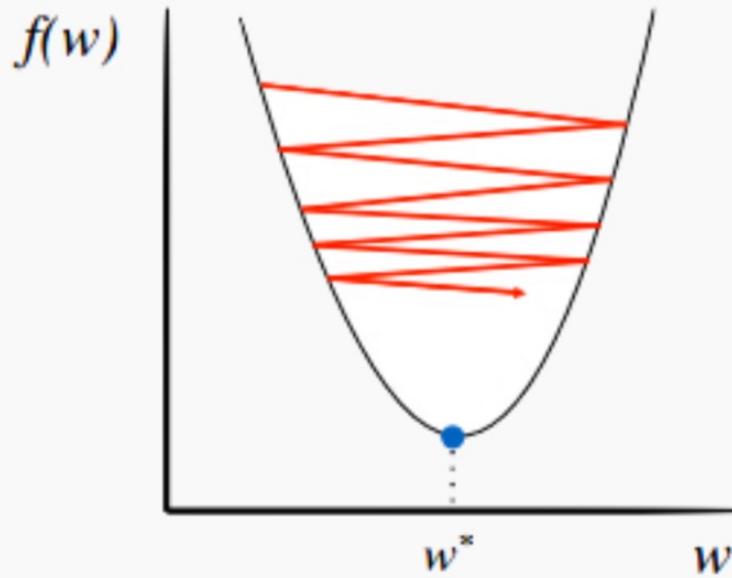
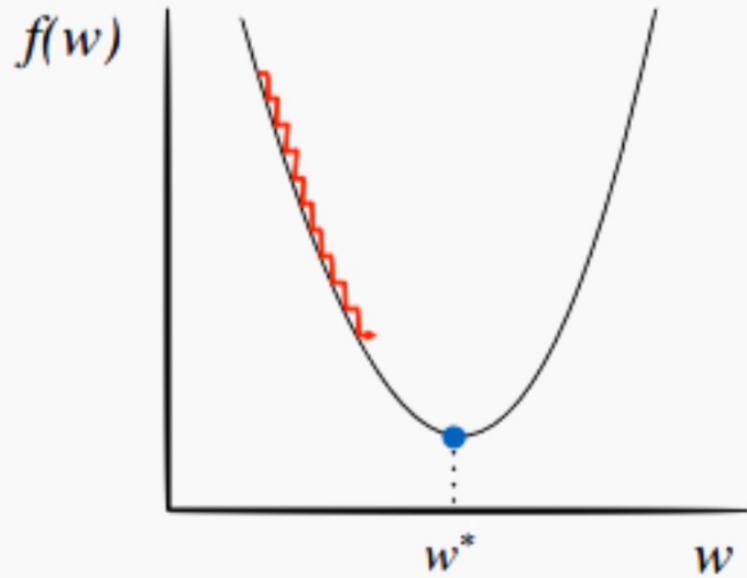
## *When do we terminate gradient descent?*

- We can limit the number of iterations in the descent. But for an arbitrary choice of maximum iterations, we cannot guarantee that we are sufficiently close to the optimum in the end.
- If the descent is stopped when the updates are sufficiently small (e.g. the residuals of  $T$  are small), we encounter a new problem: the algorithm may never terminate!

Both problems have to do with the magnitude of the learning rate,  $\lambda$ .

# Choosing a Learning Rate

For a constant learning rate,  $\lambda$ , if  $\lambda$  is too small, it takes too many iterations to reach the optimum.



If  $\lambda$  is too large, the algorithm may ‘bounce’ around the optimum and never get sufficiently close.

# Choosing a Learning Rate

Choosing  $\lambda$ :

- If  $\lambda$  is a constant, then it should be tuned through cross validation.
- For better results, use a variable  $\lambda$ . That is, let the value of  $\lambda$  depend on the gradient

$$\lambda = h(\|\nabla f(x)\|),$$

where  $\|\nabla f(x)\|$  is the magnitude of the gradient,  $\nabla f(x)$ . So

- around the optimum, when the gradient is small,  $\lambda$  should be small
- far from the optimum, when the gradient is large,  $\lambda$  should be larger

# AdaBoost

# Motivation for AdaBoost

Using the language of gradient descent also allow us to connect gradient boosting for regression to a boosting algorithm often used for classification, AdaBoost.

In classification, we *typically* want to minimize the classification error:

$$\text{Error} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}(y_n \neq \hat{y}_n), \quad \mathbb{1}(y_n \neq \hat{y}_n) = \begin{cases} 0, & y_n = \hat{y}_n \\ 1, & y_n \neq \hat{y}_n \end{cases}$$

Naively, we can try to minimize Error via gradient descent, just like we did for MSE in gradient boosting.

Unfortunately, Error is not differentiable 😞

# Motivation for AdaBoost (cont.)

**Our solution:** we replace the Error function with a differentiable function that is a good indicator of classification error.

The function we choose is called *exponential loss*

$$\text{ExpLoss} = \frac{1}{N} \sum_{n=1}^N \exp(-y_n \hat{y}_n), \quad y_n \in \{-1, 1\}$$

Exponential loss is differentiable with respect to  $\hat{y}_n$  and it is an upper bound of Error.

# Gradient Descent with Exponential Loss

We first compute the gradient for ExpLoss:

$$\nabla_{\hat{y}} \text{Exp} = [-y_1 \exp(-y_1 \hat{y}_1), \dots, -y_N \exp(-y_N \hat{y}_N)]$$

It's easier to decompose each  $y_n \exp(-y_n \hat{y}_n)$  as  $w_n y_n$ , where  $w_n = \exp(-y_n \hat{y}_n)$ .

This way, we see that the gradient is just a re-weighting applied the target values

$$\nabla \text{Exp} = [-w_1 y_1, \dots, -w_N y_N]$$

Notice that when  $y_n = \hat{y}_n$ , the weight  $w_n$  is small; when  $y_n \neq \hat{y}_n$ , the weight is larger.

# Gradient Descent with Exponential Loss

The update step in the gradient descent is

$$\hat{y}_n \leftarrow \hat{y}_n + \lambda w_n y_n, \quad n = 1, \dots, N$$

Just like in gradient boosting, we approximate the gradient,  $\lambda w_n y_n$  with a simple model,  $T^{(i)}$ , that depends on  $x_n$ .

This means training  $T^{(i)}$  on a re-weighted set of target values,

$$\{(x_1, w_1 y_1), \dots, (x_N, w_N y_N)\}$$

That is, gradient descent with exponential loss means iteratively training simple models that ***focuses on the points misclassified by the previous model.***

# AdaBoost

With a minor adjustment to the exponential loss function, we have the algorithm for gradient descent:

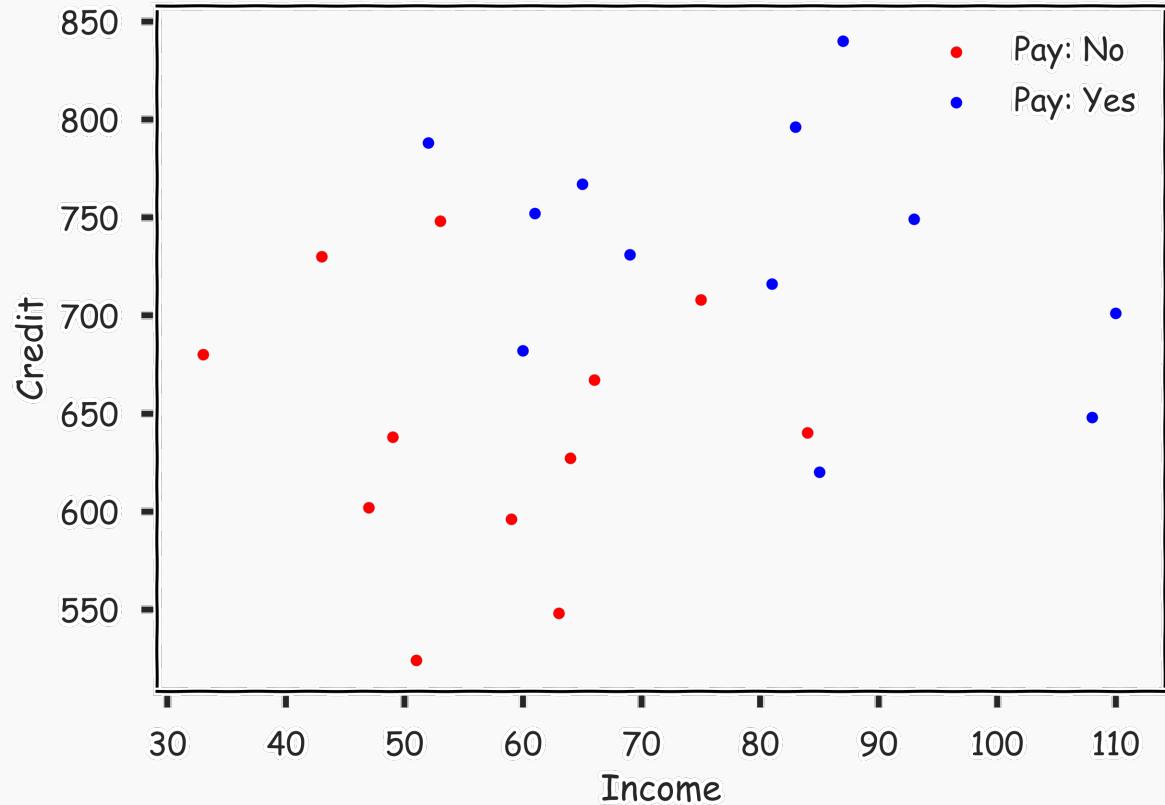
1. Choose an initial distribution over the training data,  $w_n = 1/N$ .
2. At the  $i^{th}$  step, fit a simple classifier  $T^{(i)}$  on weighted training data  $\{(x_1, w_1 y_1), \dots, (x_N, w_N y_N)\}$
3. Update the weights:

$$w_n \leftarrow \frac{w_n \exp(-\lambda^{(i)} y_n T^{(i)}(x_n))}{Z}$$

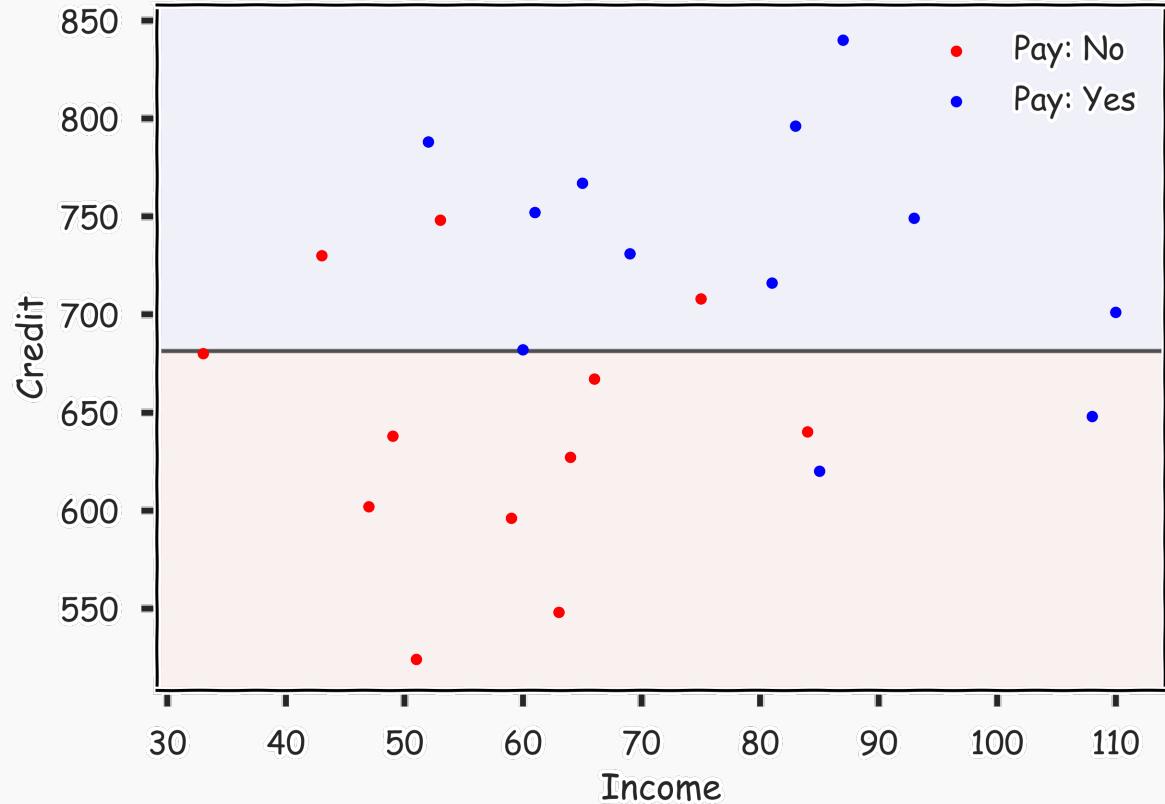
where  $Z$  is the normalizing constant for the collection of updated weights

4. Update  $T: T \leftarrow T + \lambda^{(i)} T^{(i)}$   
where  $\lambda$  is the learning rate.

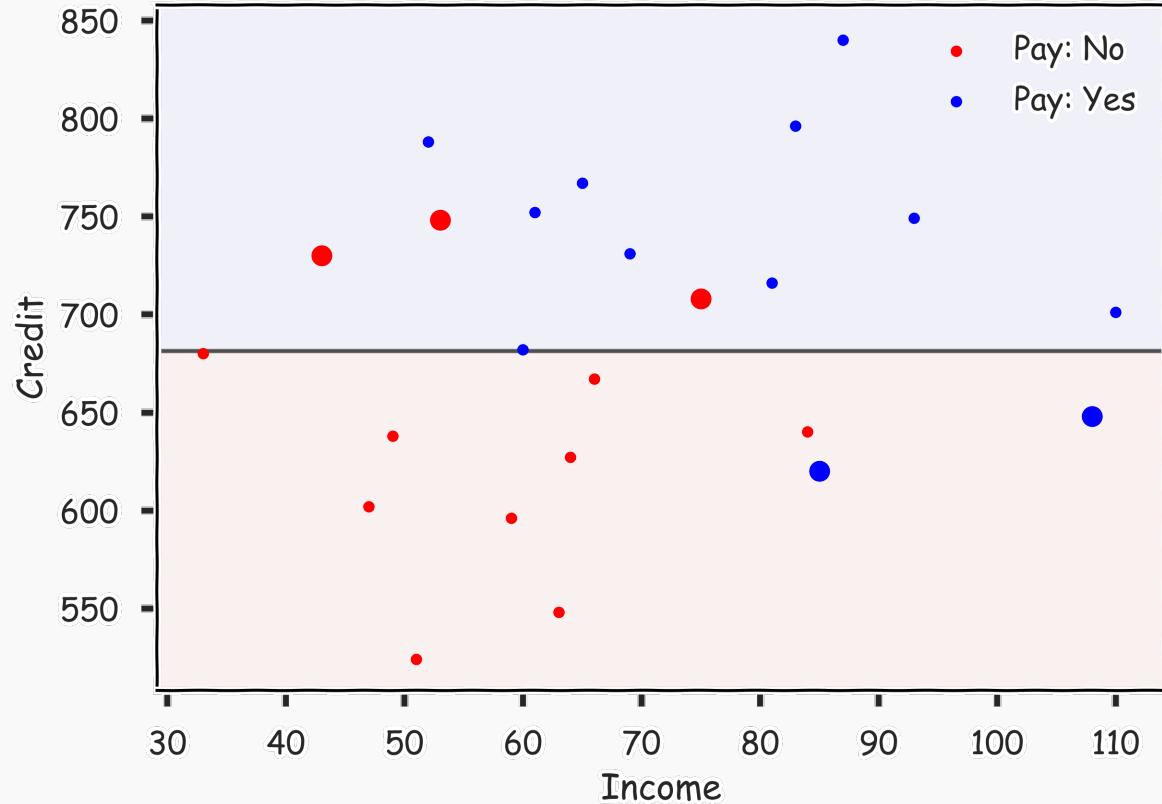
# AdaBoost: start with equal weights



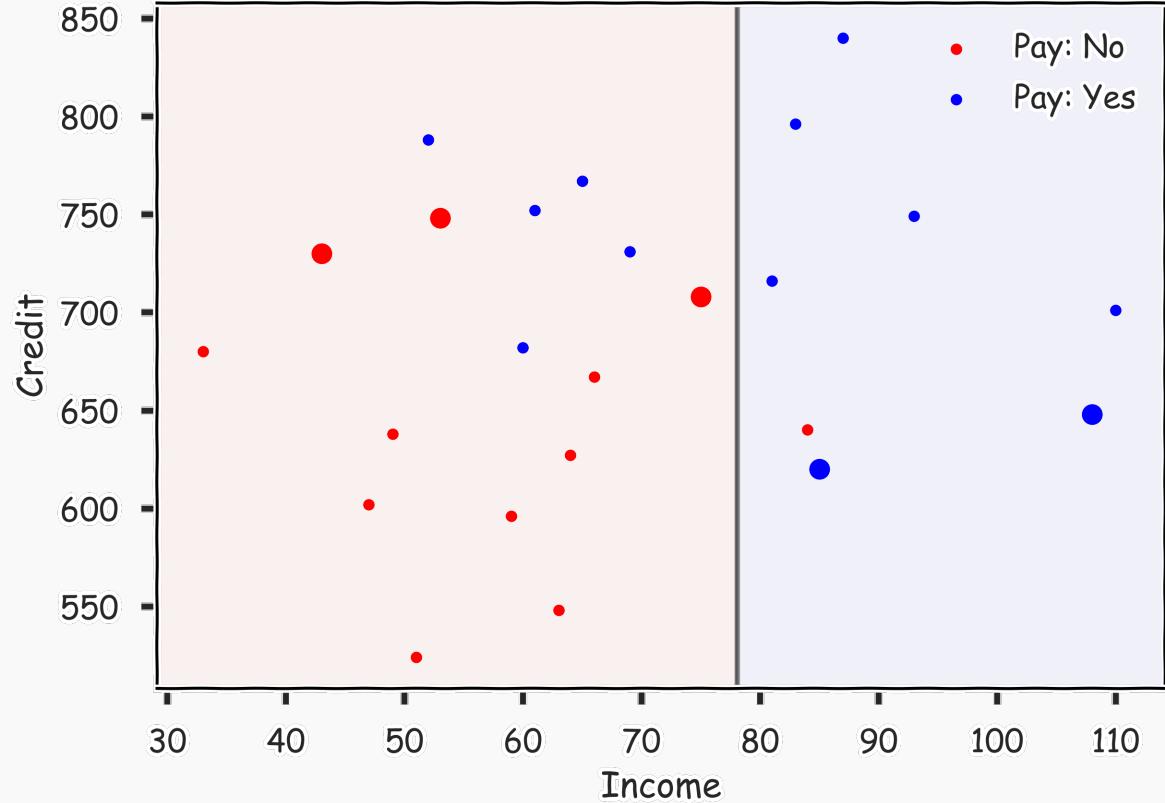
# AdaBoost: fit a simple decision tree



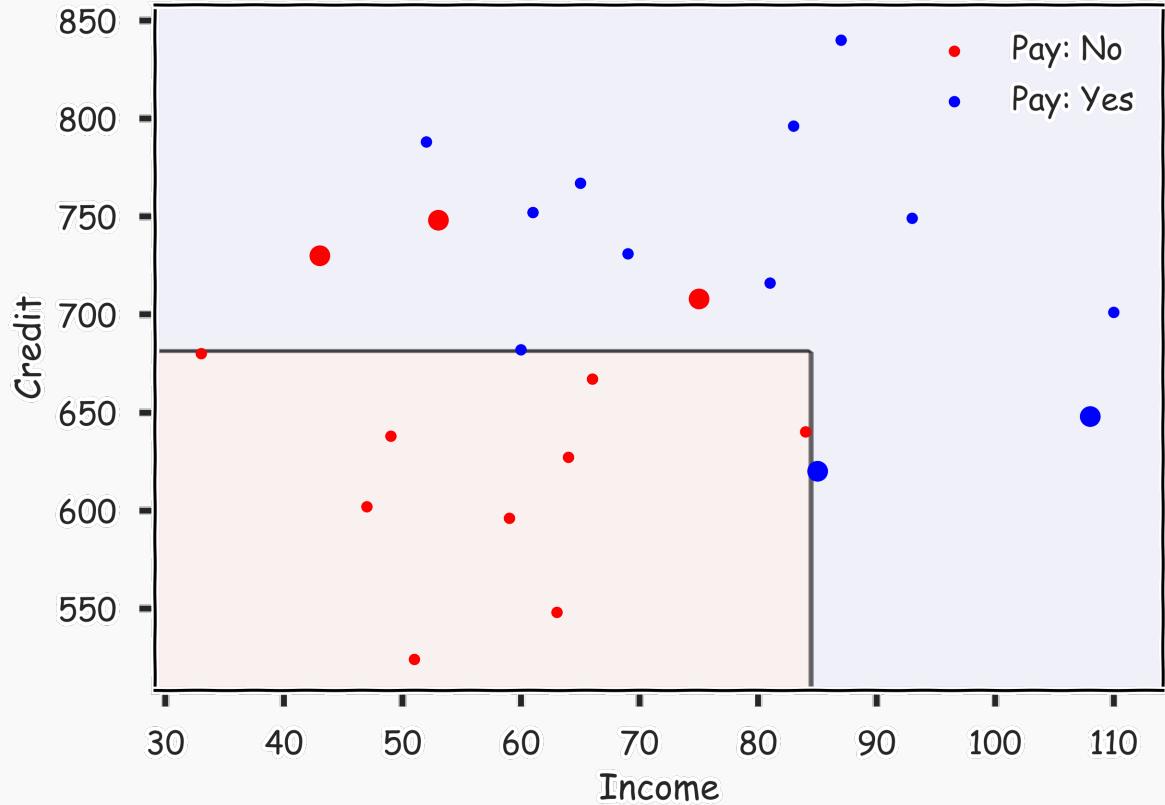
# AdaBoost: update the weights (errors are weighted higher)



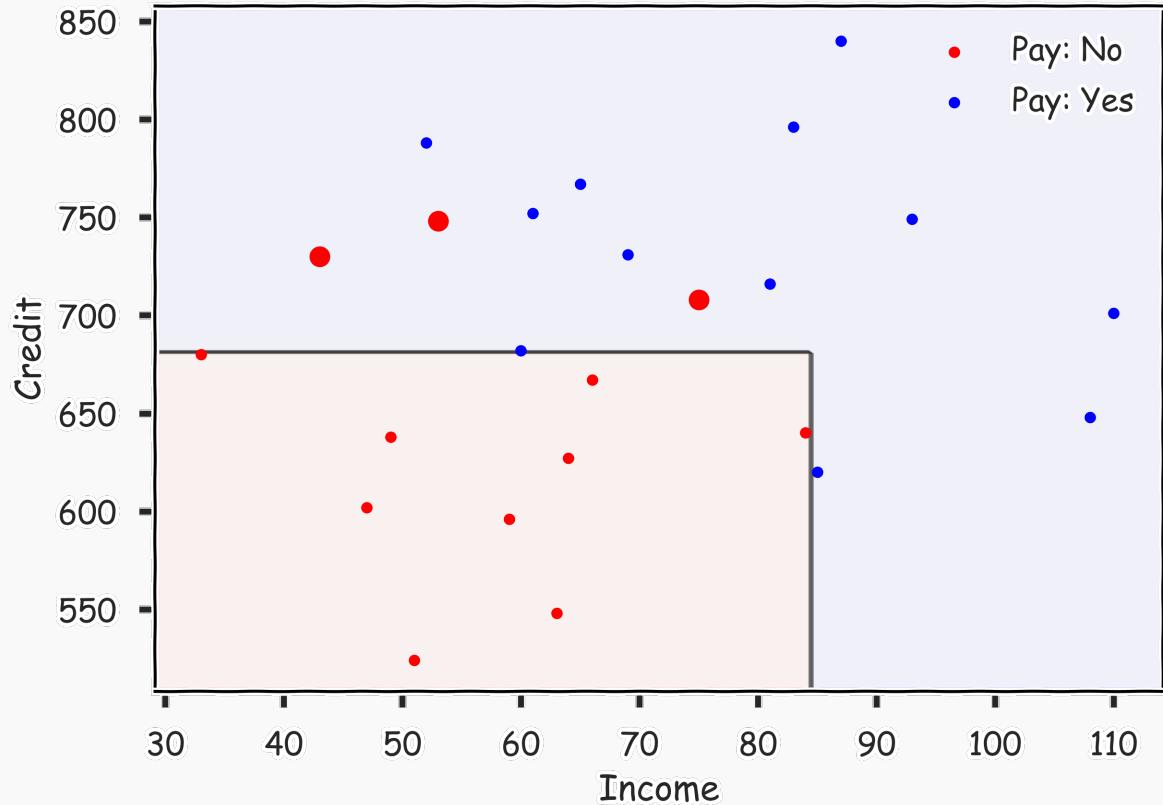
# AdaBoost: fit another simple decision tree on re-weighted data



# AdaBoost: combine the trees

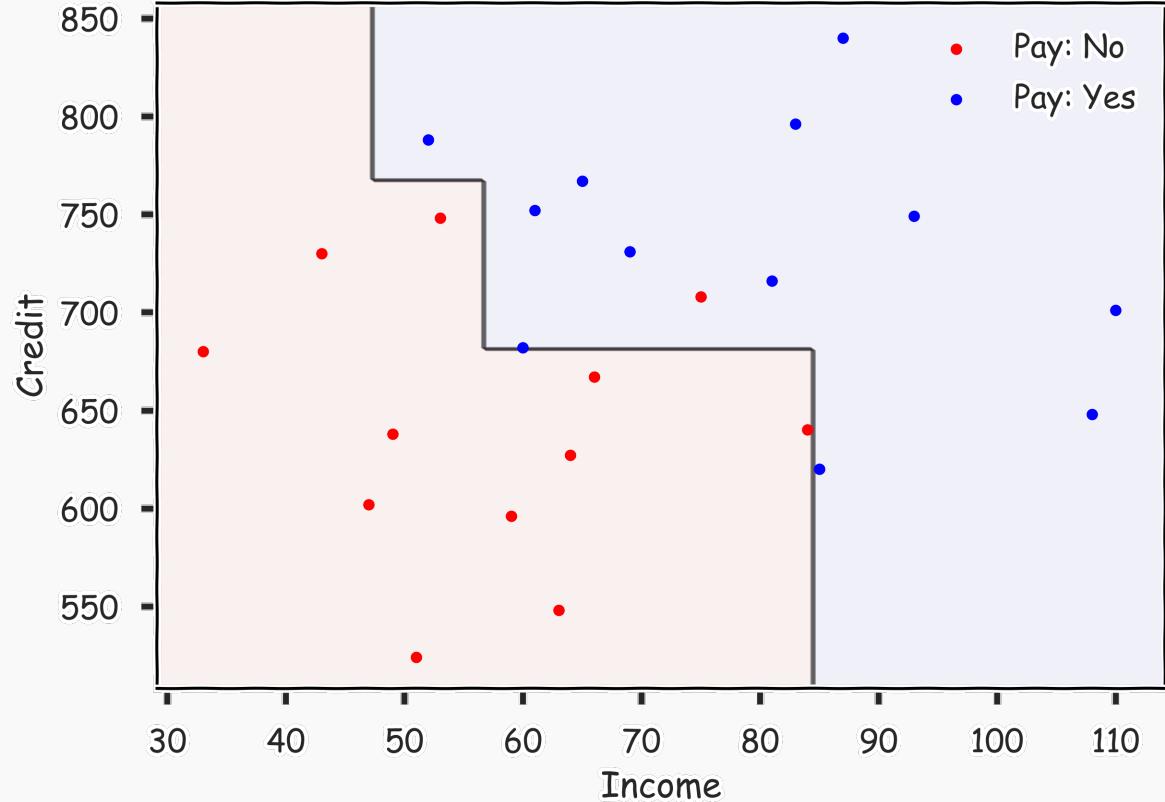


# AdaBoost: update the weights



$$w_n \leftarrow \frac{w_n \exp(-\lambda^{(i)} y_n T^{(i)}(x_n))}{Z}$$

# AdaBoost: fit a third, simple decision tree on re-weighted data



$$T \leftarrow T + \lambda^{(i)} T^{(i)}$$

# Choosing the Learning Rate

Unlike in the case of gradient boosting for regression, we can analytically solve for the optimal learning rate for AdaBoost, by optimizing:

$$\operatorname{argmin}_{\lambda} \frac{1}{N} \sum_{n=1}^N \exp \left[ -y_n (T + \lambda^{(i)} T^{(i)}(x_n)) \right]$$

Doing so, we get that

$$\lambda^{(i)} = \frac{1}{2} \ln \frac{1-\epsilon}{\epsilon}, \quad \epsilon = \sum_{n=1}^N w_n \mathbb{1}(y_n \neq T^{(i)}(x_n))$$

# Final thoughts on Boosting

There are few implementations on boosting:

- XGBoost: An efficient Gradient Boosting Decision
- LGBM: Light Gradient Boosted Machines. It is a library for training GBMs developed by Microsoft, and it competes with XGBoost
- CatBoost: A new library for Gradient Boosting Decision Trees, offering appropriate handling of categorical features

# Final thoughts on Boosting

Increasing the number of *trees* can lead to overfitting.

**Question:** Why?

# Summary of Tree Ensemble Methods

In **bagging**, the trees are grown independently on random samples of the observations. Consequently, the trees tend to be quite similar to each other. Thus, bagging can get caught in local optima and can fail to thoroughly explore the model space.

In **random forests**, the trees are once again grown independently on random samples of the observations. However, each split on each tree is performed using a random subset of the features, thereby decorrelating the trees, and leading to a more thorough exploration of model space relative to bagging.

In **boosting**, we only use the original data, and do not draw any random samples. The trees are grown successively, using a “slow” learning approach: each new tree is fit to the signal that is left over from the earlier trees, and shrunken down before it is used.

## Question:

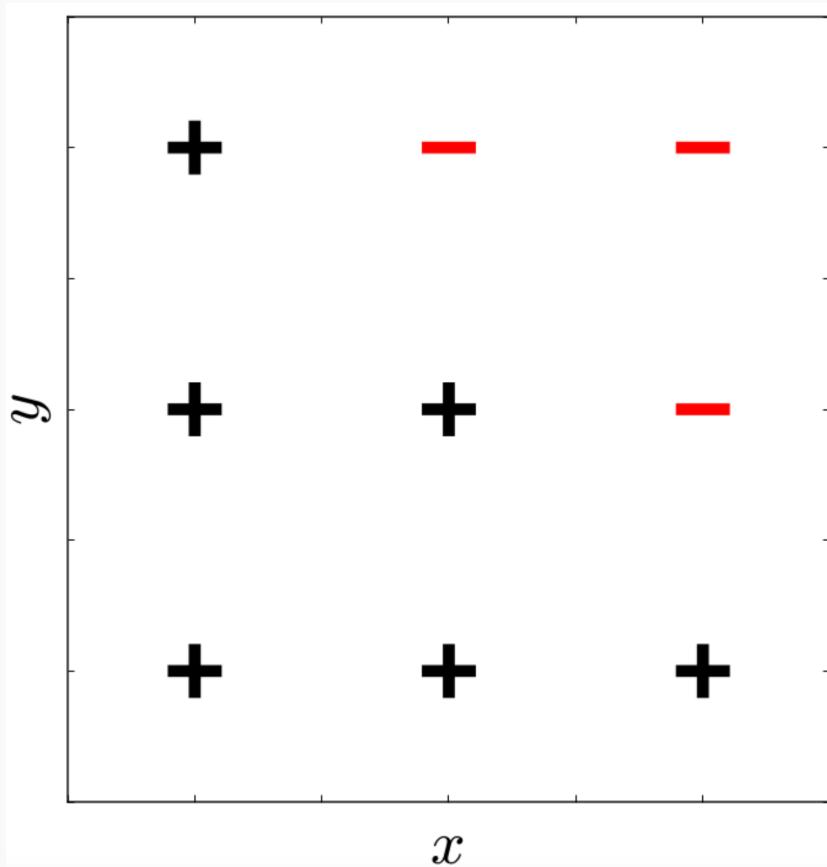
Recall that AdaBoost learns a classifier  $H$  using a weighted sum of weak learners  $h_t$  as follows

$$H(x) = \text{sgn} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$$

In this question we will use decision trees as our weak learners, which classify a point as  $\{1, -1\}$  based on a sequence of threshold splits on its features (here  $x, y$ ).

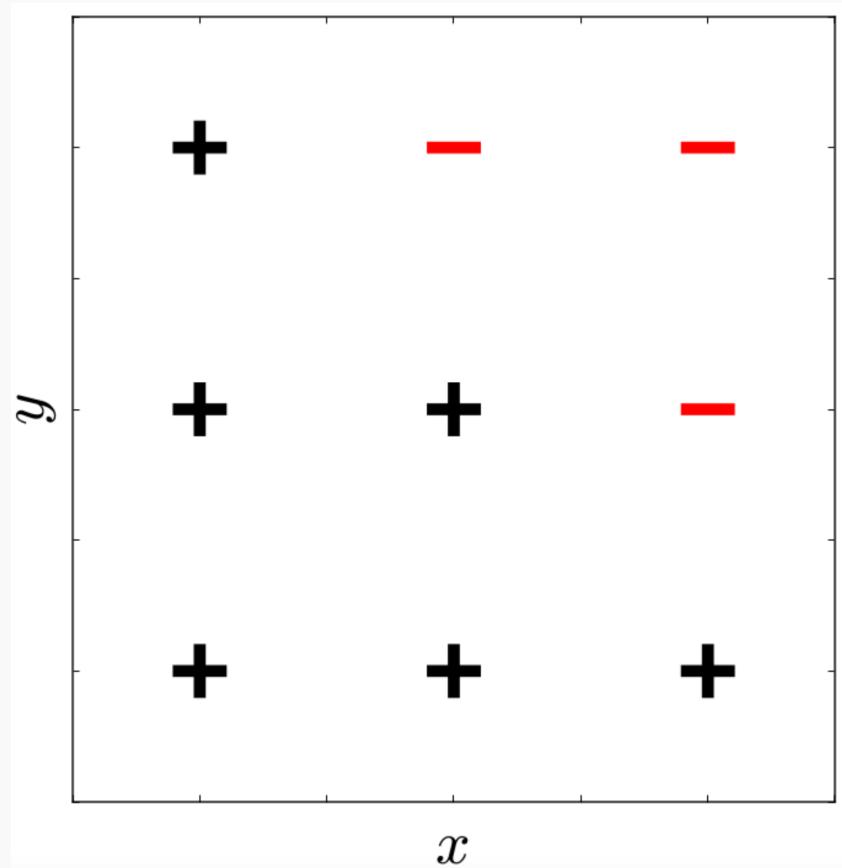
# Question:

- (a) Assume that our weak learners are decision trees of depth 1 (i.e. decision stumps), which minimize the weighted training error.
- (i) Using the dataset below, draw the decision boundary learned by  $h_1$ .



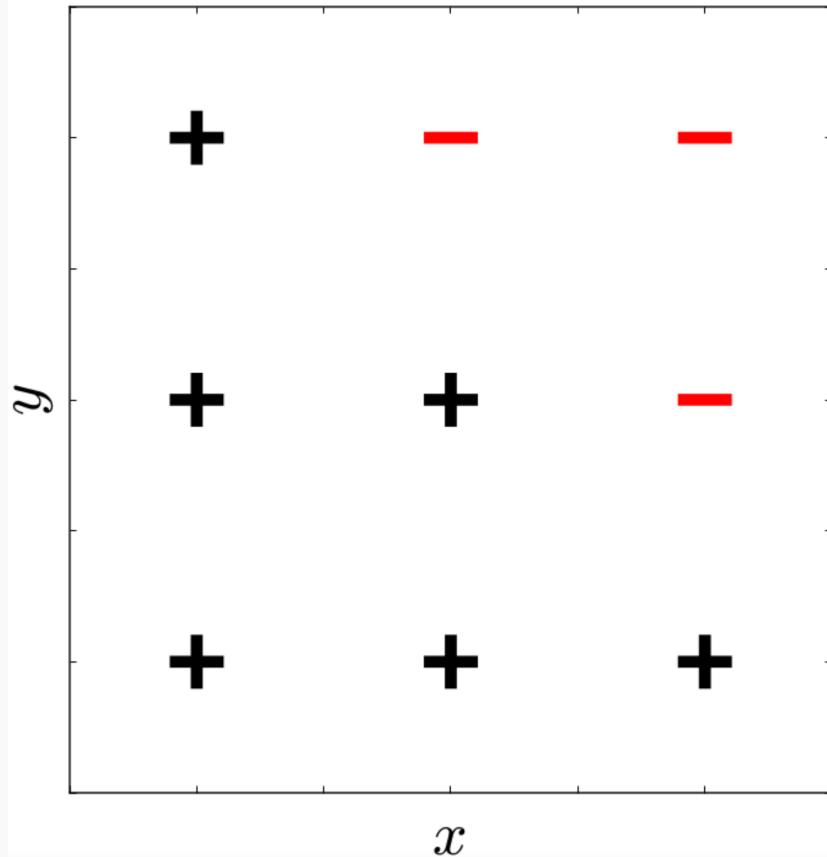
## Question:

(ii) On the dataset below, circle the point(s) with the highest weights on the second iteration, and draw the decision boundary learned by  $h_2$ . Explain the reasons.



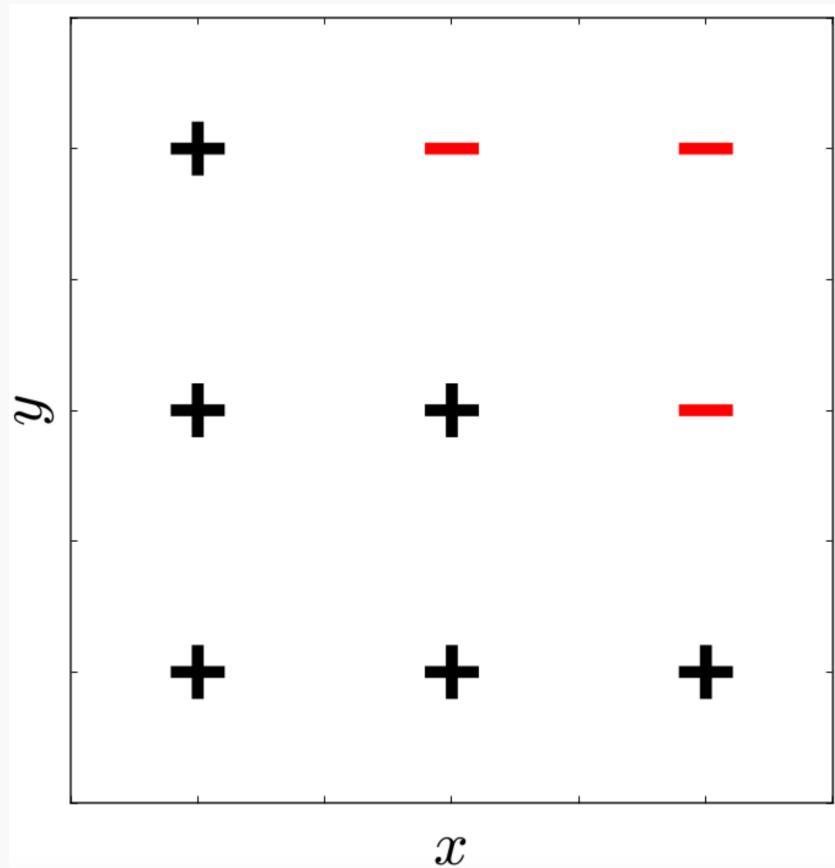
# Question:

- (iii) On the dataset below, draw the decision boundary of  $H = \text{sgn}(\alpha_1 h_1 + \alpha_2 h_2)$ . Explain the reasons.  
(Hint, you do not need to explicitly compute the  $\alpha$ 's).



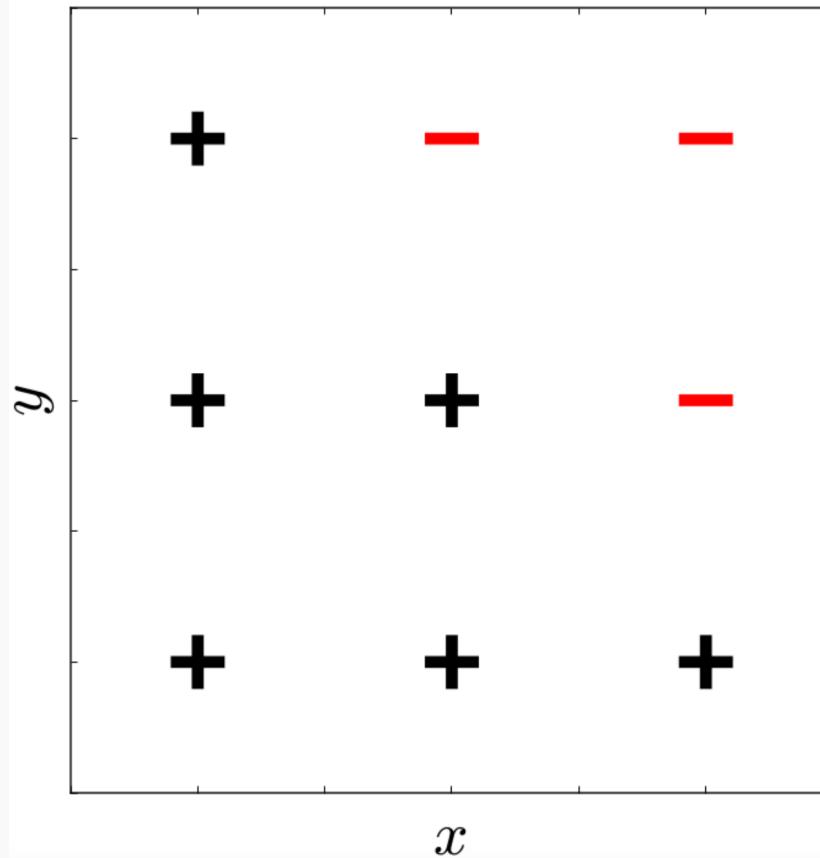
## Question:

- (b) Now assume that our weak learners are decision trees of maximum depth 2, which minimize the weighted training error.
- (i) Using the dataset below, draw the decision boundary learned by  $h_1$ .



# Question:

- (ii) On the dataset below, circle the point(s) with the highest weights on the second iteration, and draw the decision boundary learned by  $h_2$ . Explain the reasons.



## Question:

(iii) On the dataset below, draw the decision boundary of  $H = \text{sgn}(\alpha_1 h_1 + \alpha_2 h_2)$ . Explain the reasons. (Hint, you do not need to explicitly compute the  $\alpha$ 's).

