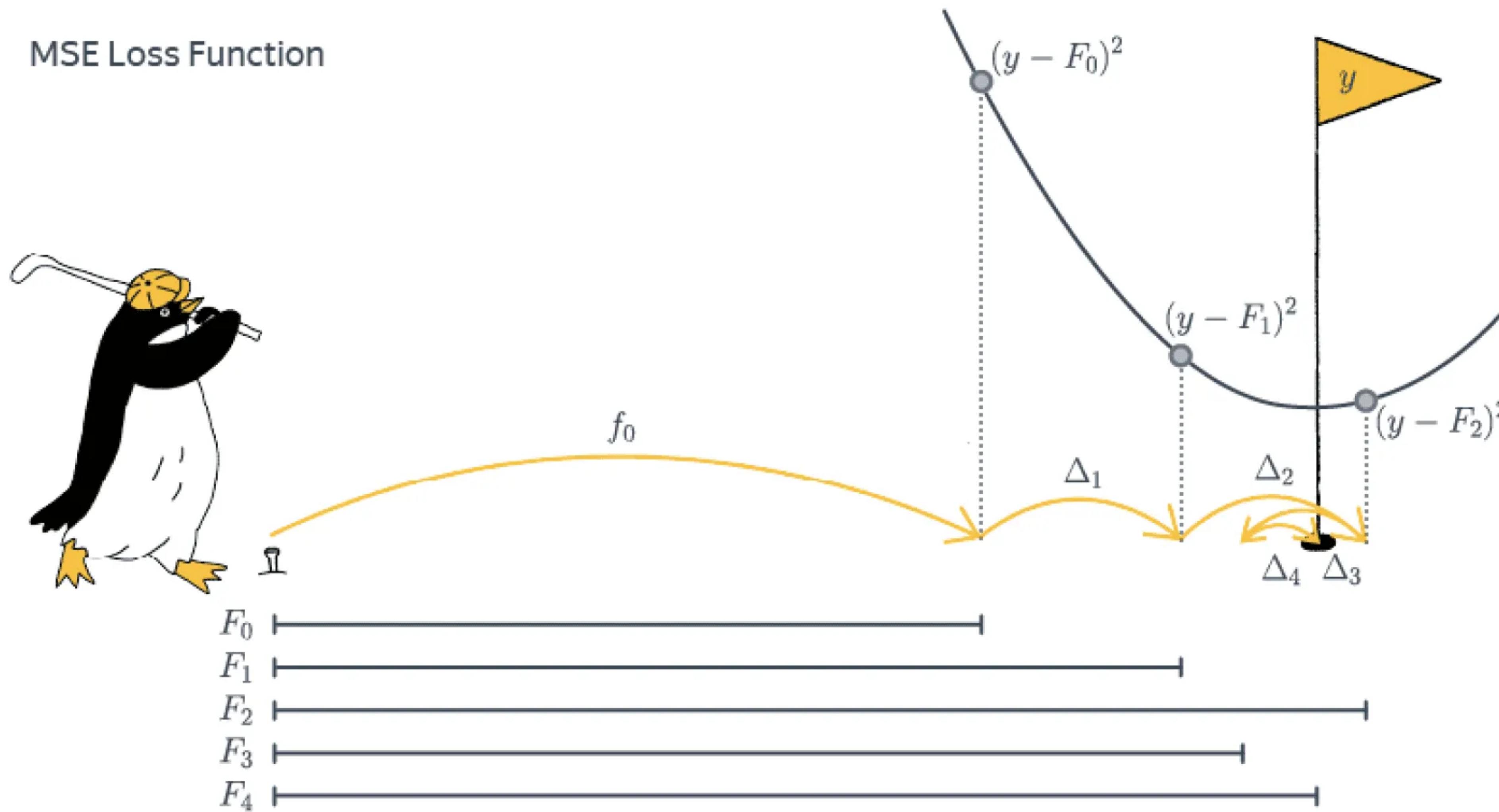


# Ensembles

# Gradient Boosting



# GB/ идея

- Возьмём простые базовые модели
- Будем строить композицию последовательно и жадно
- Каждая следующая модель будет строиться так, чтобы максимально корректировать ошибки построенных моделей

$$a_N(x) = \sum_{n=1}^N b_n(x)$$

- Обучение первой модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, b_1(x_i)) \rightarrow \min_{b_1(x)}$$

# GB/ идея

$$a_N(x) = \sum_{n=1}^N b_n(x)$$

- Обучение первой модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, b_1(x_i)) \rightarrow \min_{b_1(x)}$$

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow \min_{b_N(x)}$$

# **GB**/первые проблемы

- В бустинге базовые модели обучаются последовательно
- Каждая следующая корректирует ошибки уже построенных
- В общем случае получается функционал, на который может быть сложно обучать деревья

# GB/MSE как функция потерь

$$a_N(x) = \sum_{n=1}^N b_n(x)$$

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a_{N-1}(x_i) + b_N(x_i) - y_i)^2 \rightarrow \min_{b_N(x)}$$

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_N(x_i) - \underbrace{(y_i - a_{N-1}(x_i))}_{{s_i}^{(N)}} \right)^2 \rightarrow \min_{b_N(x)}$$

# **GB/ итоговый план**

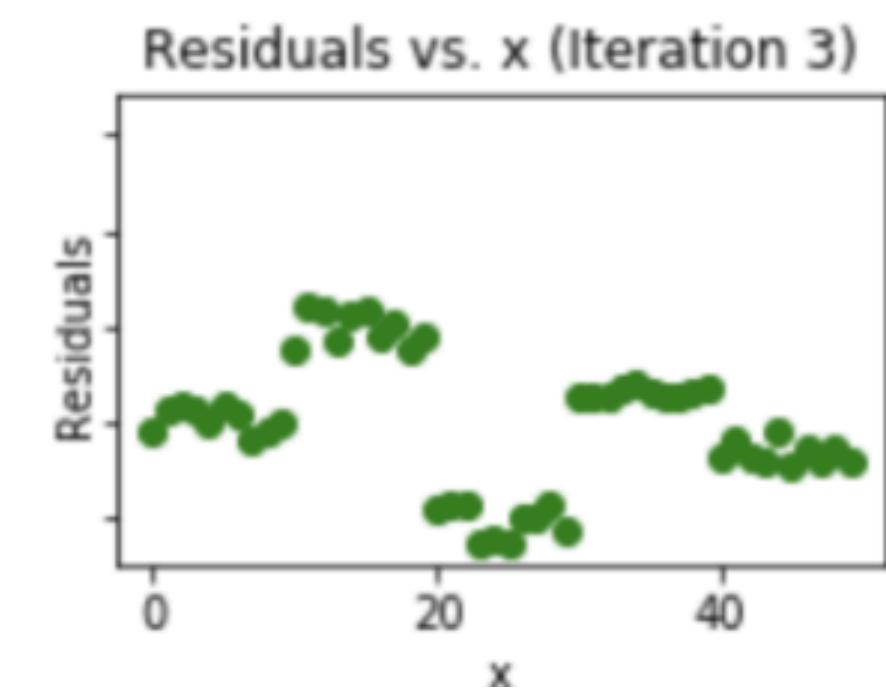
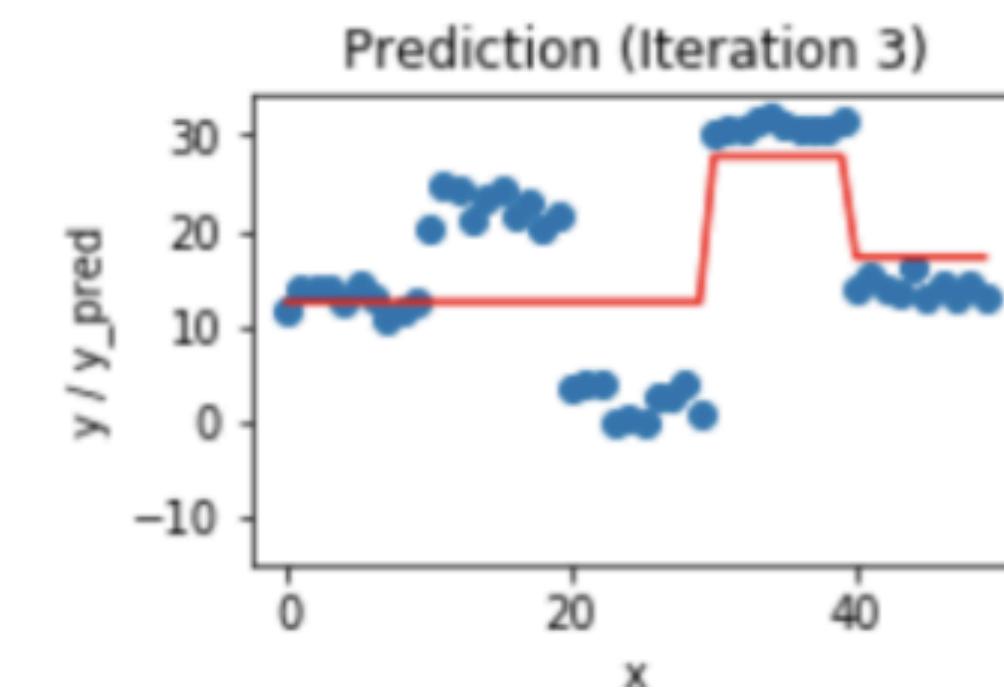
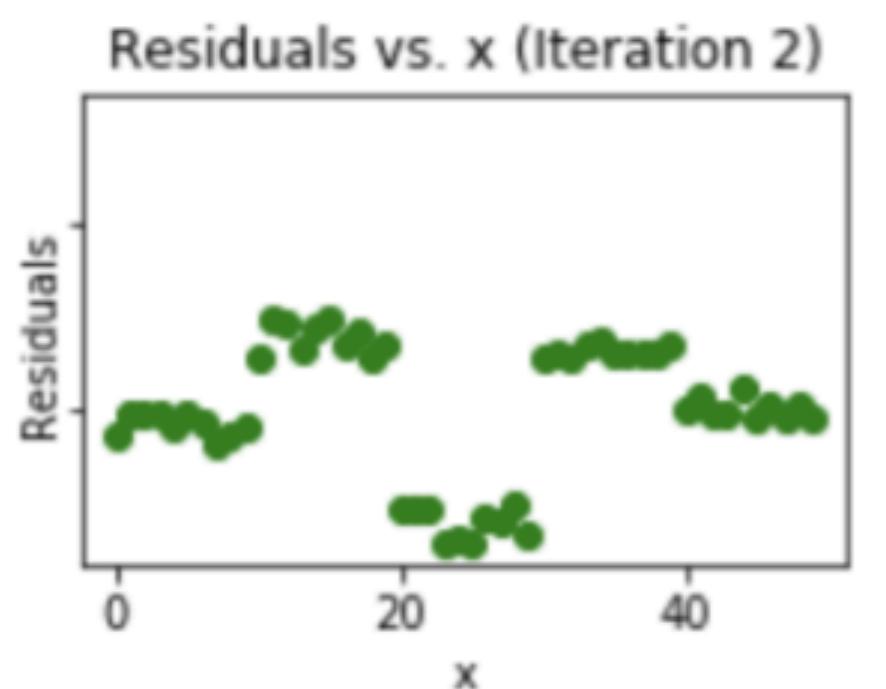
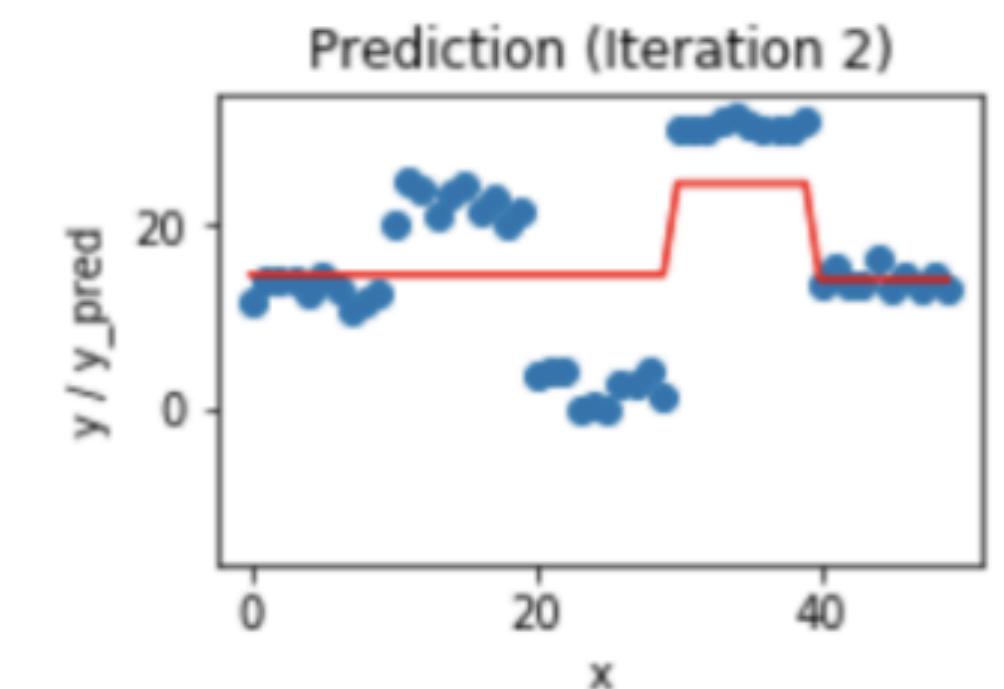
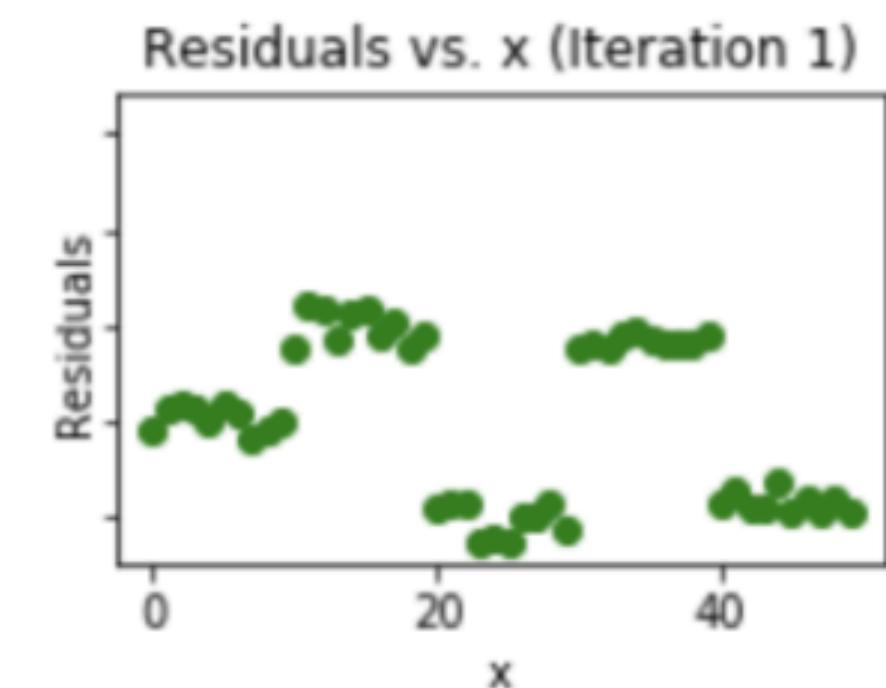
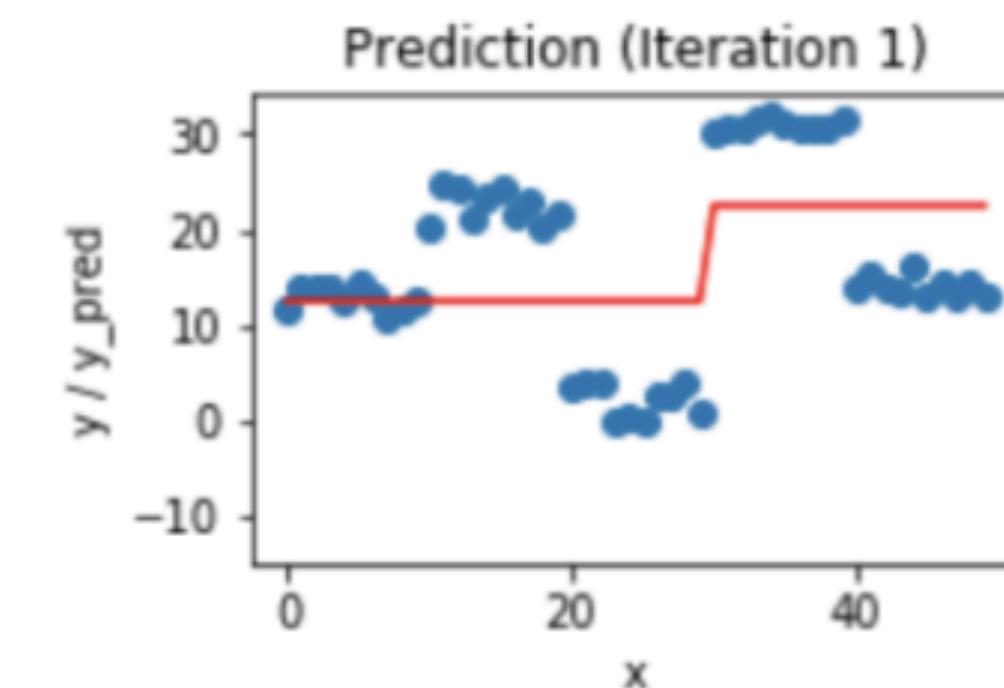
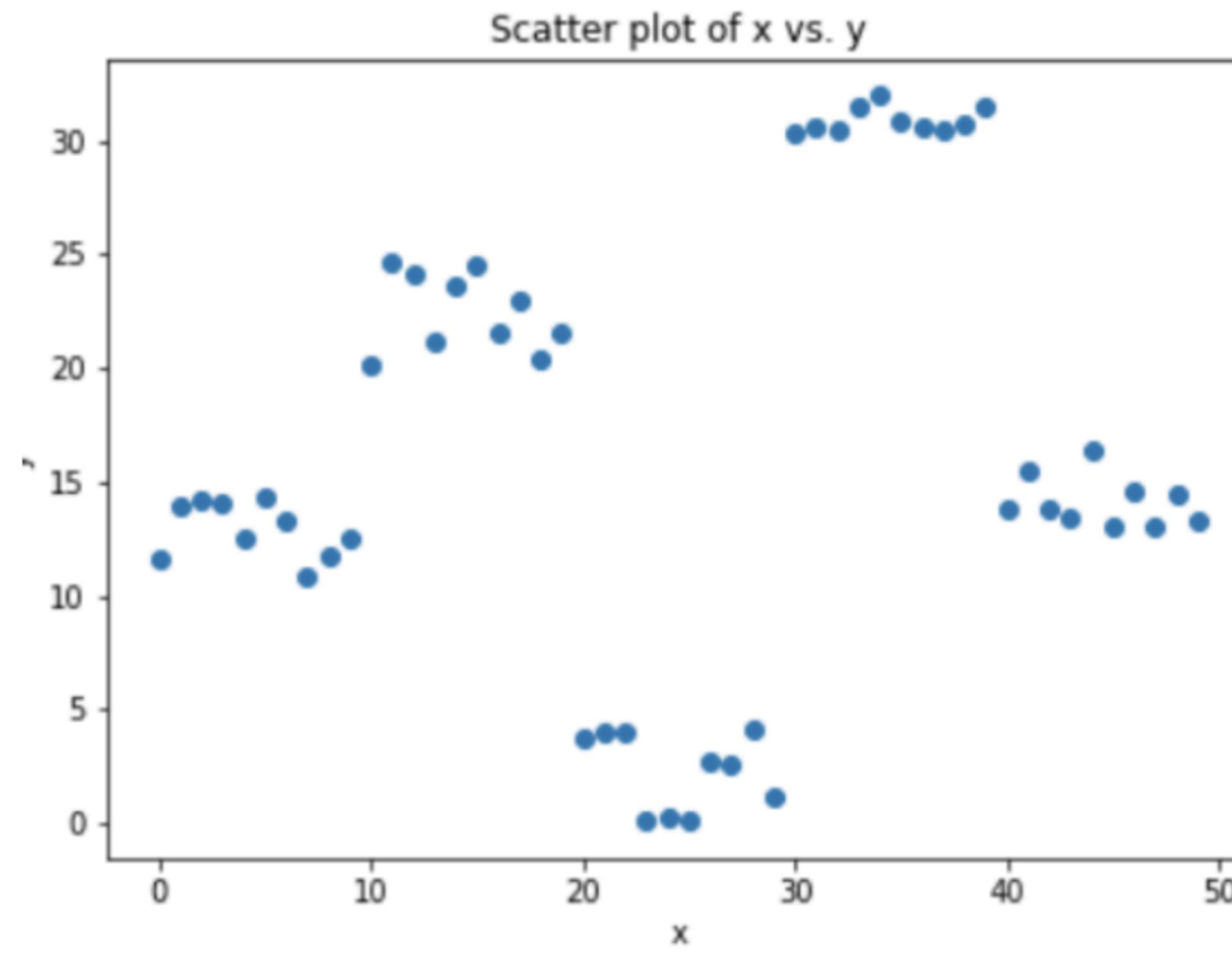
$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_N(x_i) - s_i^{(N)} \right)^2 \rightarrow \min_{b_N(x)}$$

- $s_i^{(N)} = y_i - a_{N-1}(x_i)$  — остатки

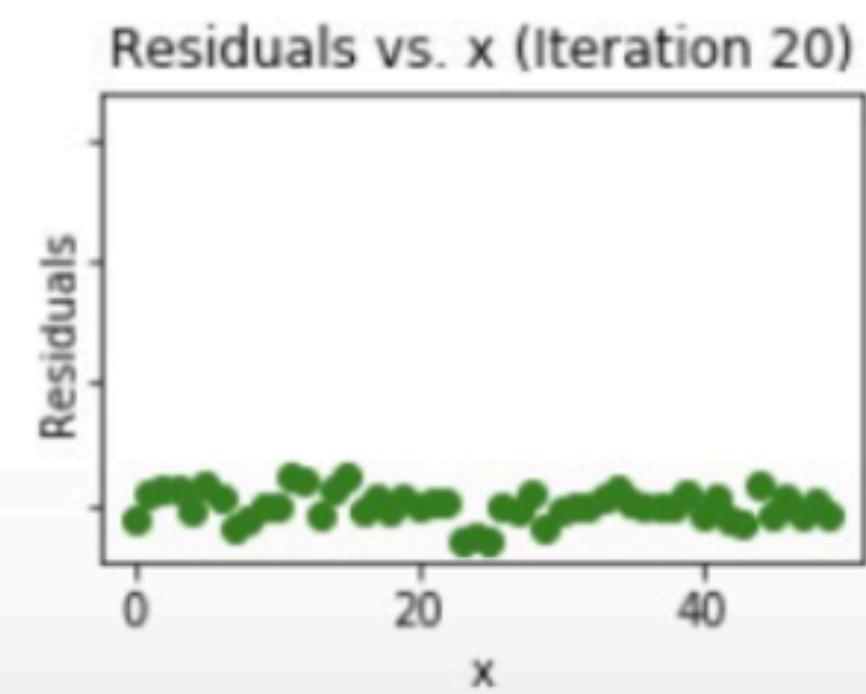
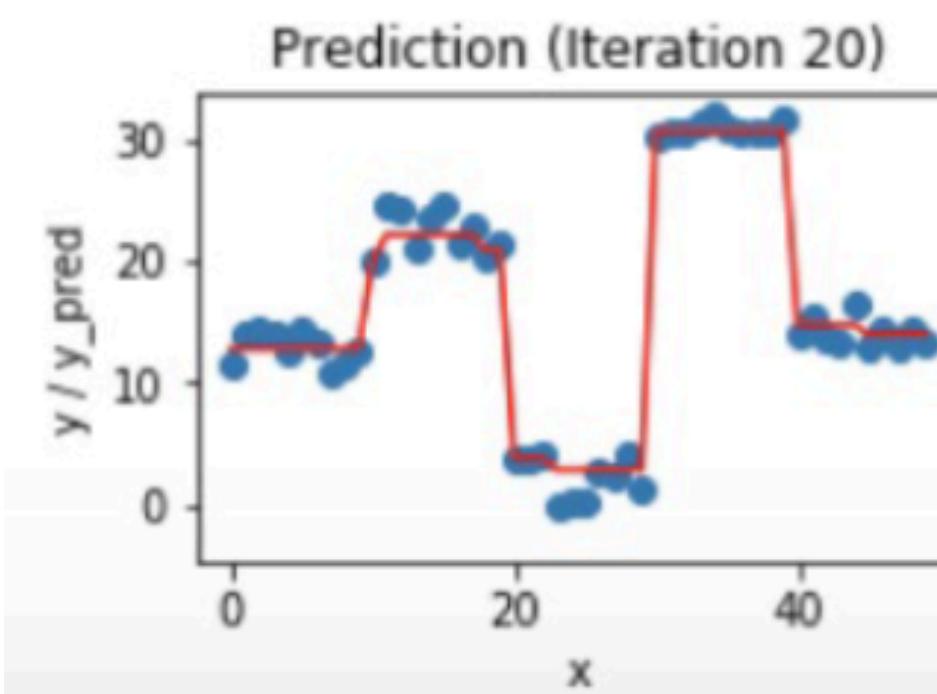
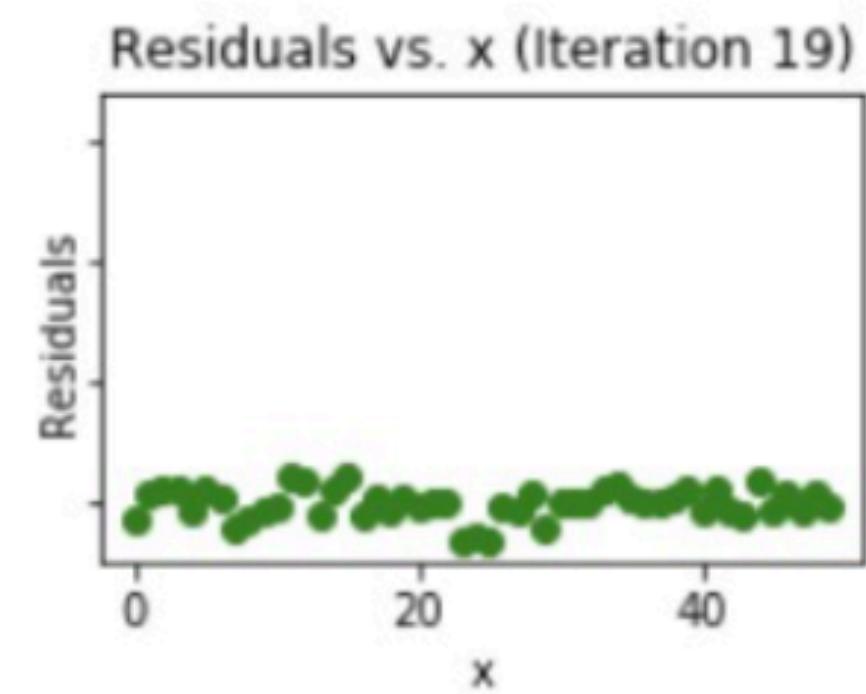
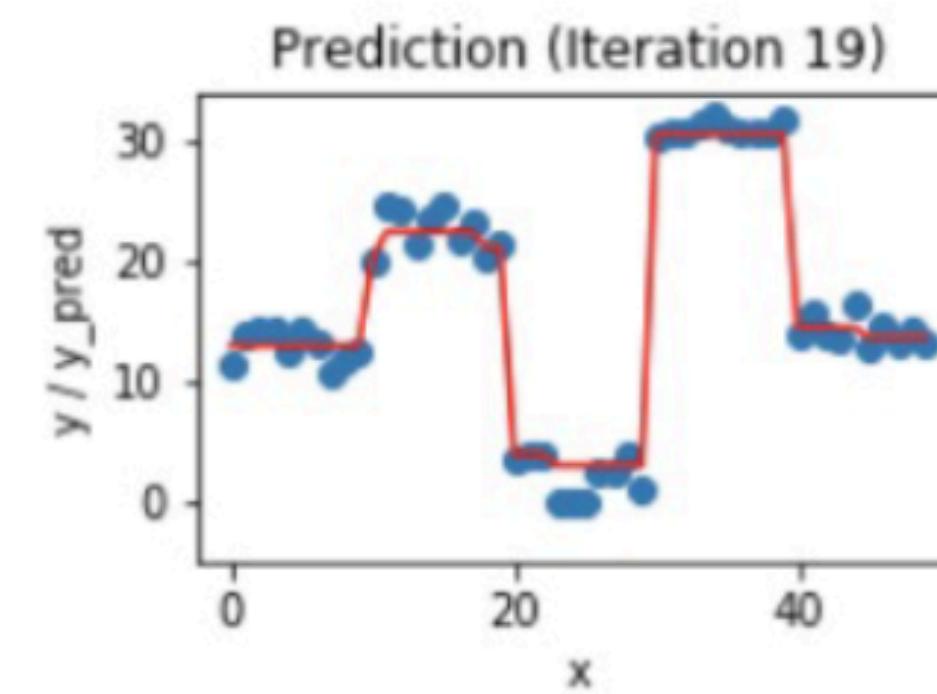
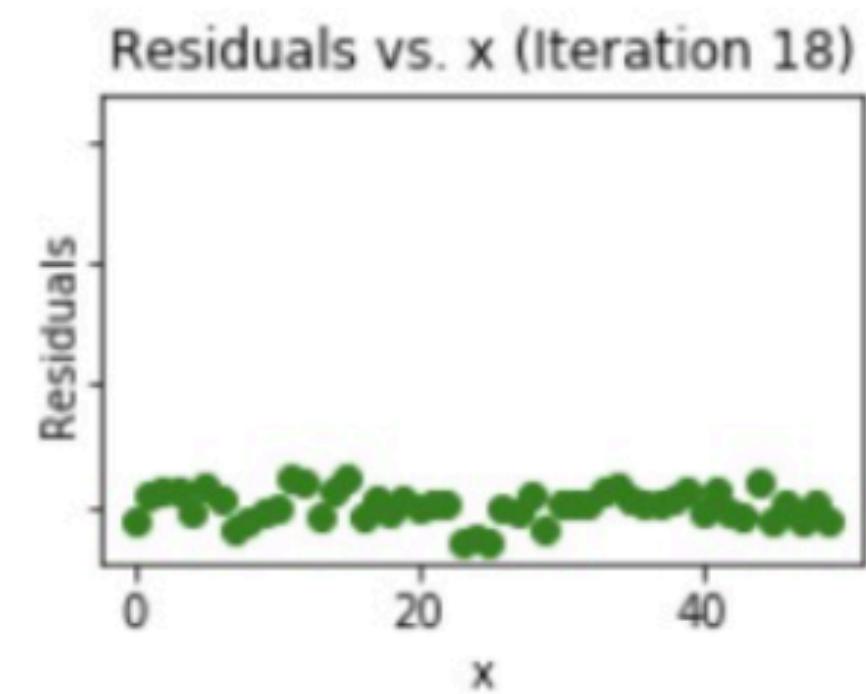
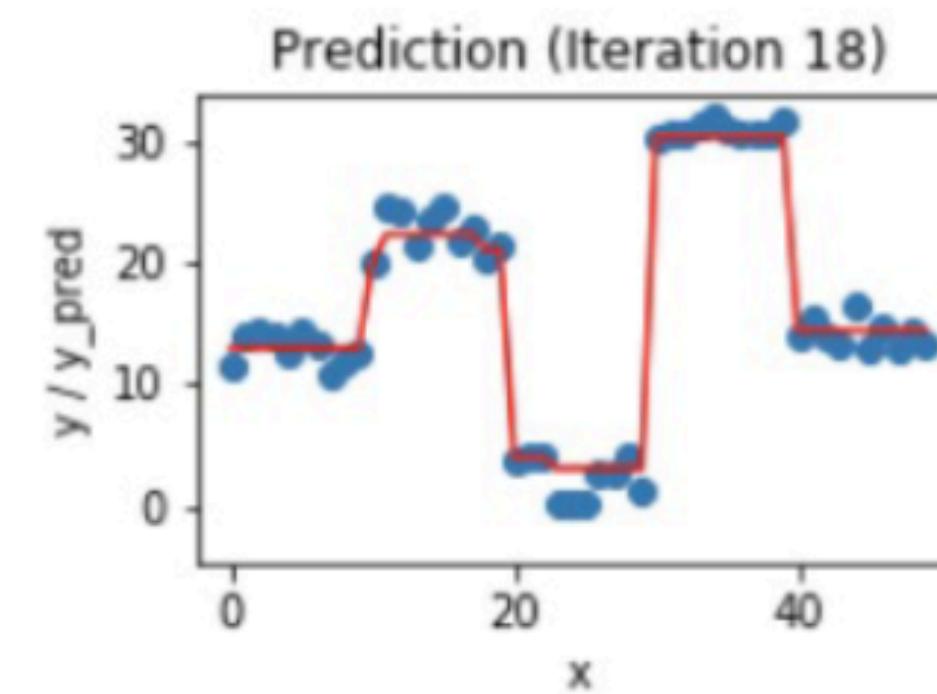
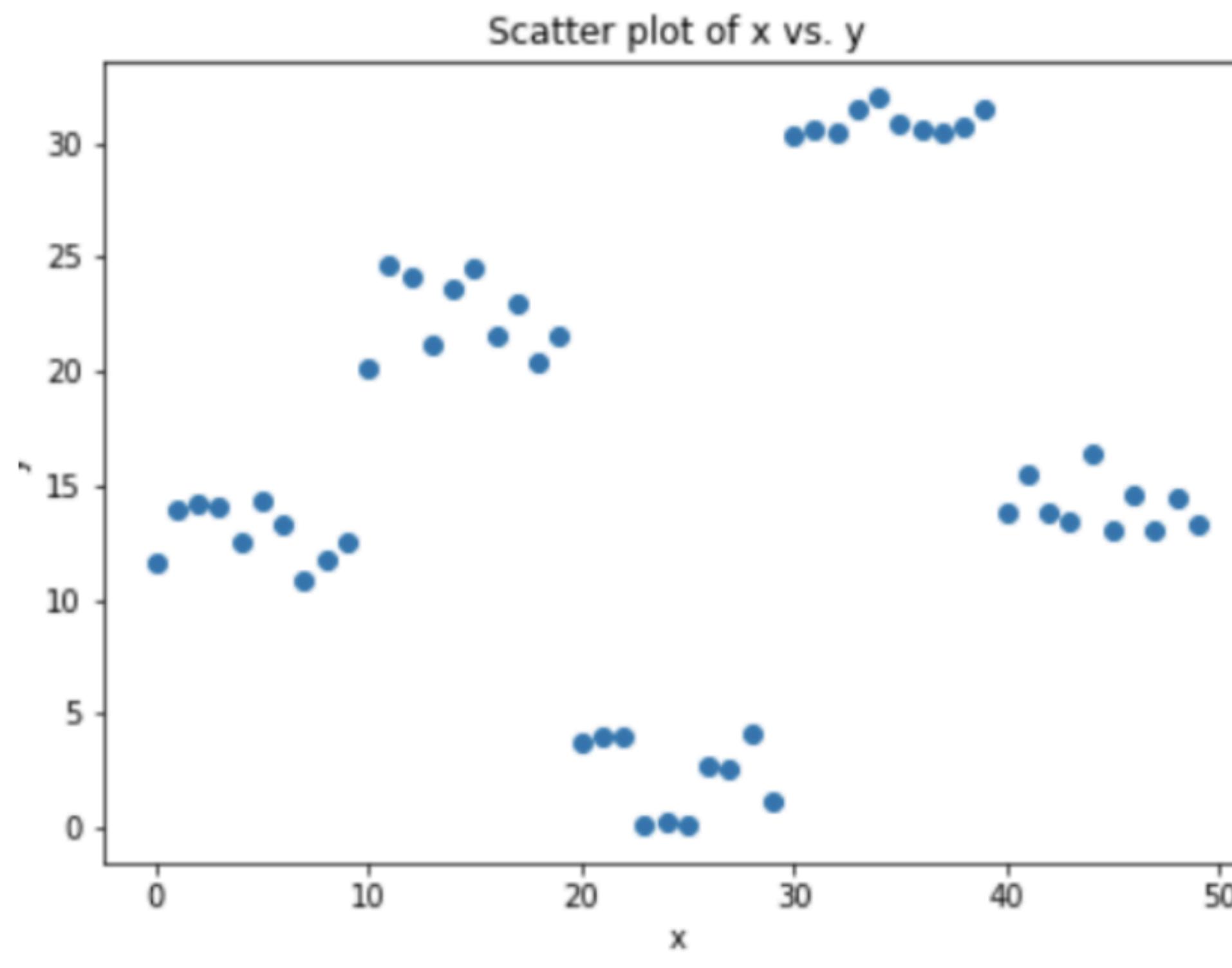
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (b_1(x_i) - y_i)^2 \rightarrow \min_{b_1(x)}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_2(x_i) - (y_i - b_1(x_i)) \right)^2 \rightarrow \min_{b_2(x)}$$

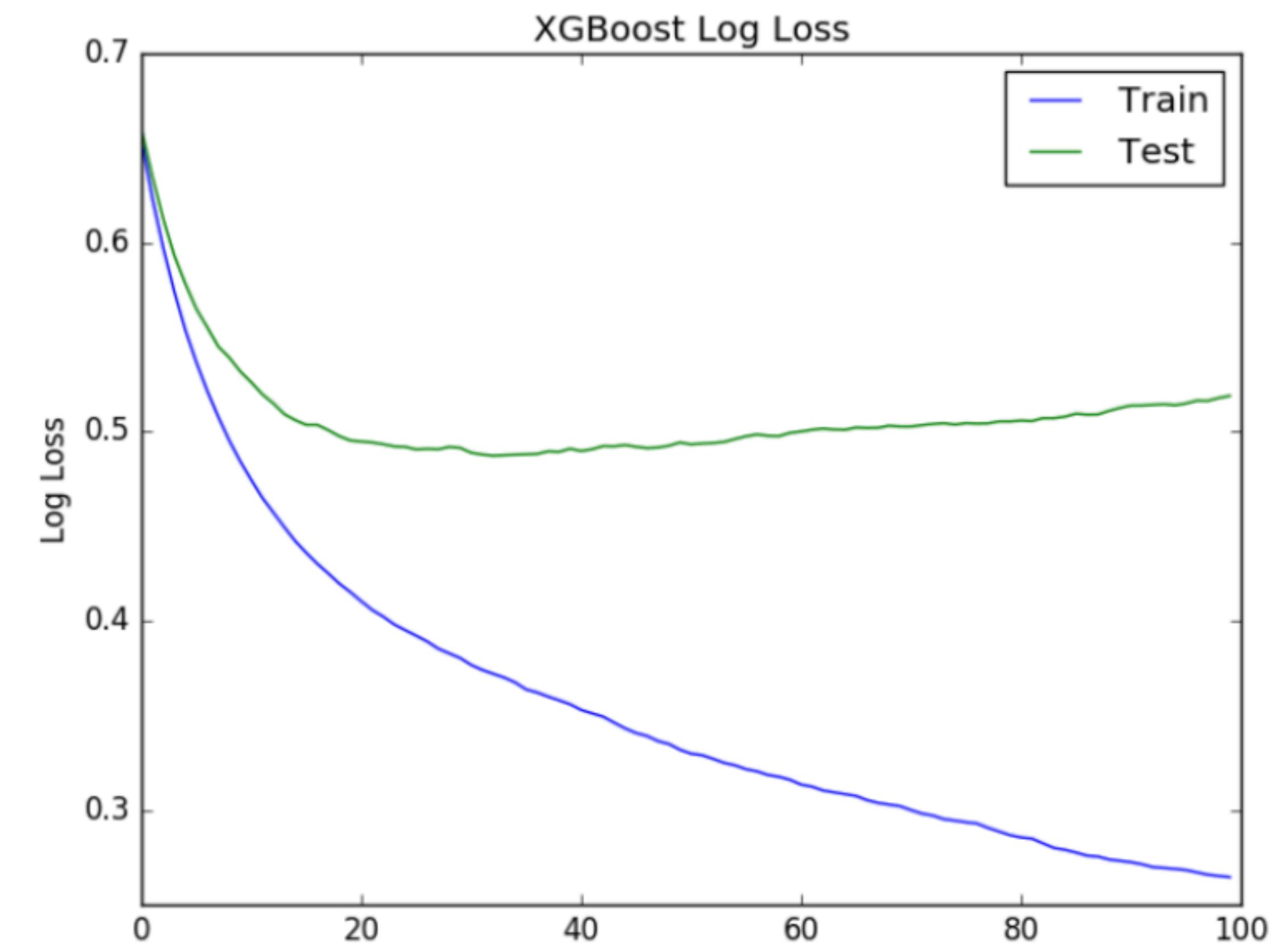
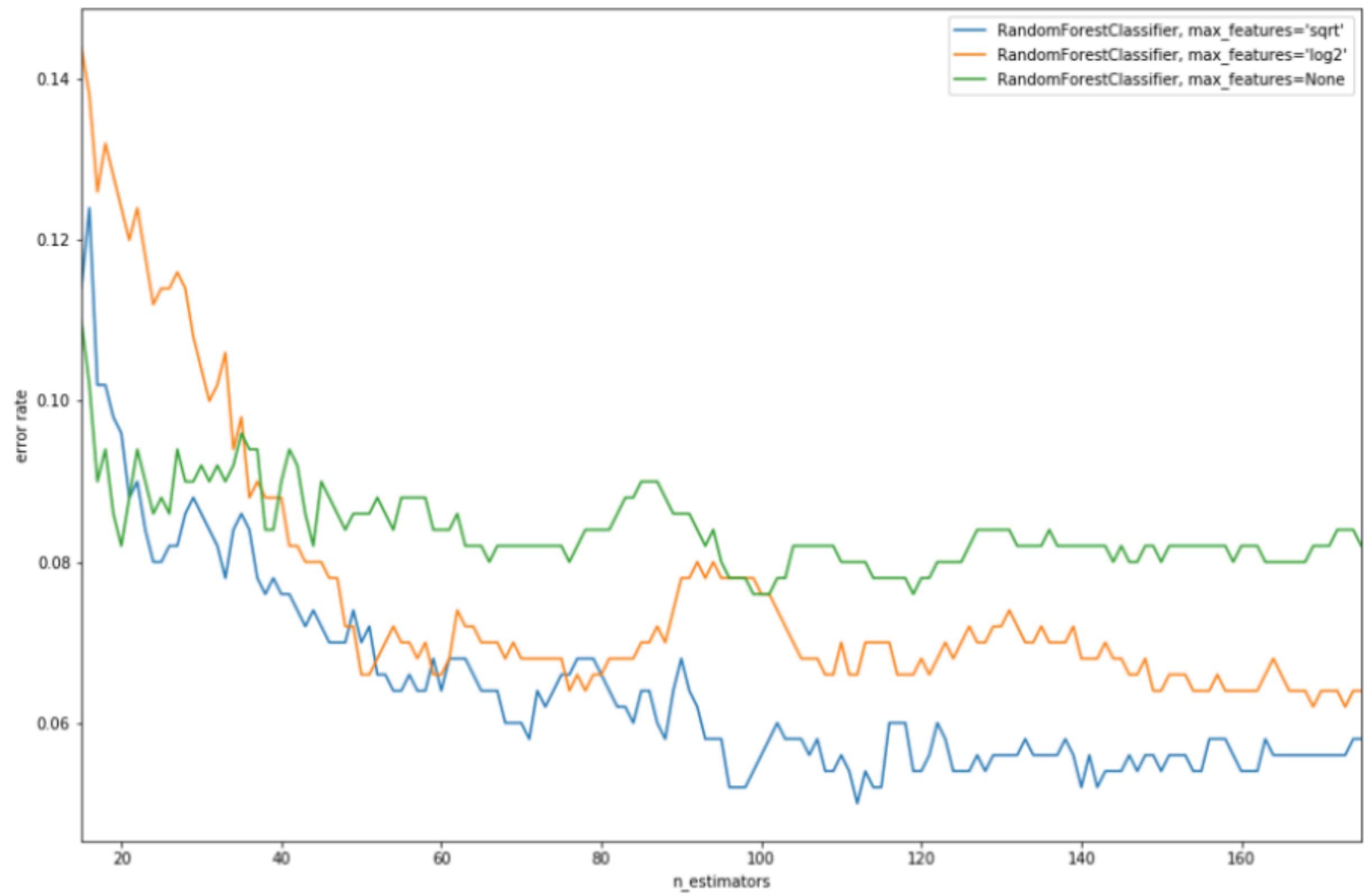
# GB/пример обучения



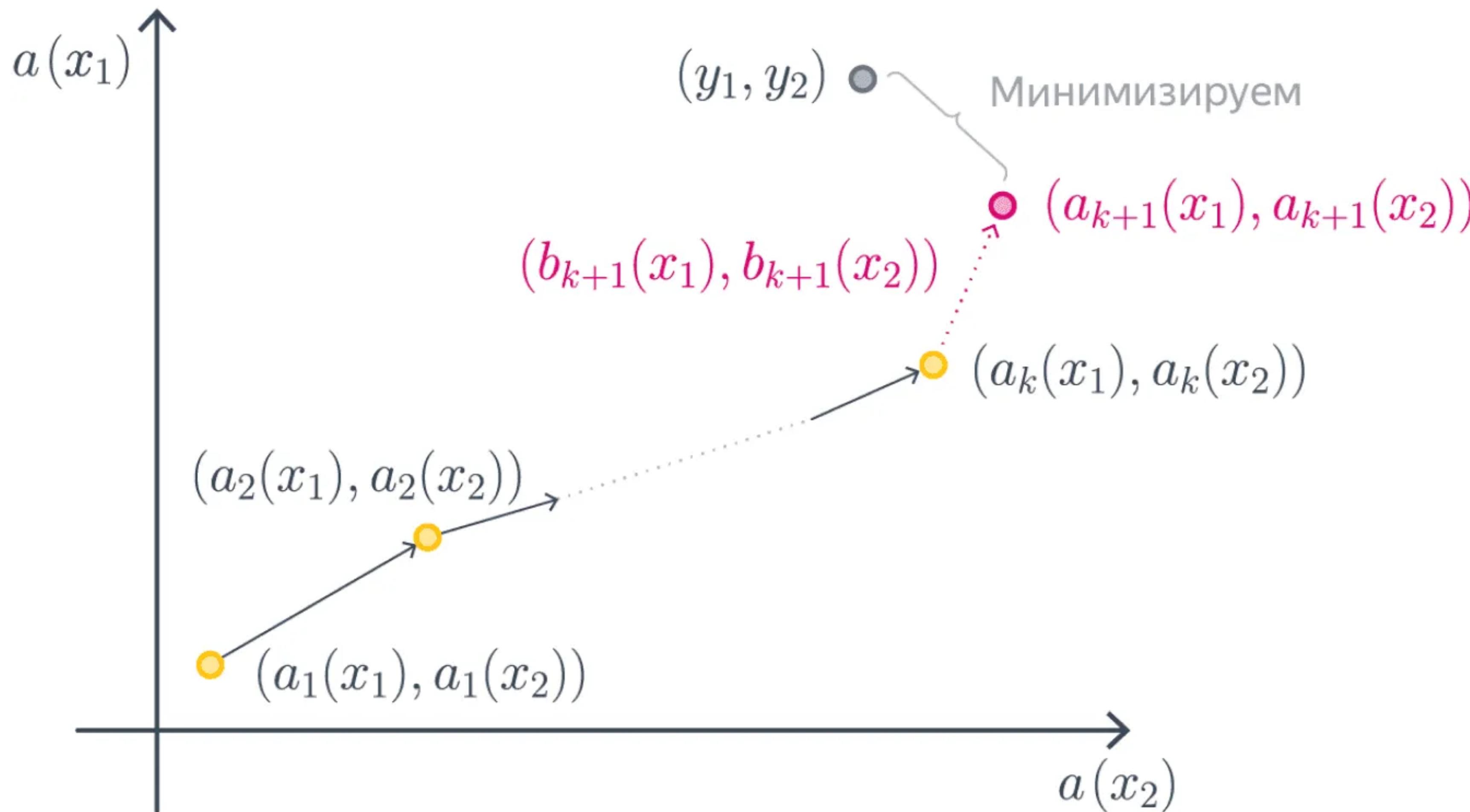
# GB/пример обучения



# GB/переобучение



# Gradient Boosting w/o MSE



# GB / логистическая регрессия

$$a_N(x) = \text{sign} \sum_{n=1}^N b_n(x)$$

$$L(y, z) = \log(1 + \exp(-yz))$$

- Может, просто обучаться на остатки, как в MSE?

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log \left( 1 + \exp \left( - (y_i - a_{N-1}(x_i)) b_N(x_i) \right) \right) \rightarrow \min_{b_N(x)}$$

- Если  $y_i = a_{N-1}(x_i)$ , то объект не участвует в обучении
- Иначе  $y_i - a_{N-1}(x_i) = \pm 2$

# GB/ логистическая регрессия

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log \left( 1 + \exp \left( - \frac{y_i - a_{N-1}(x_i)}{2} b_N(x_i) \right) \right) \rightarrow \min_{b_N(x)}$$

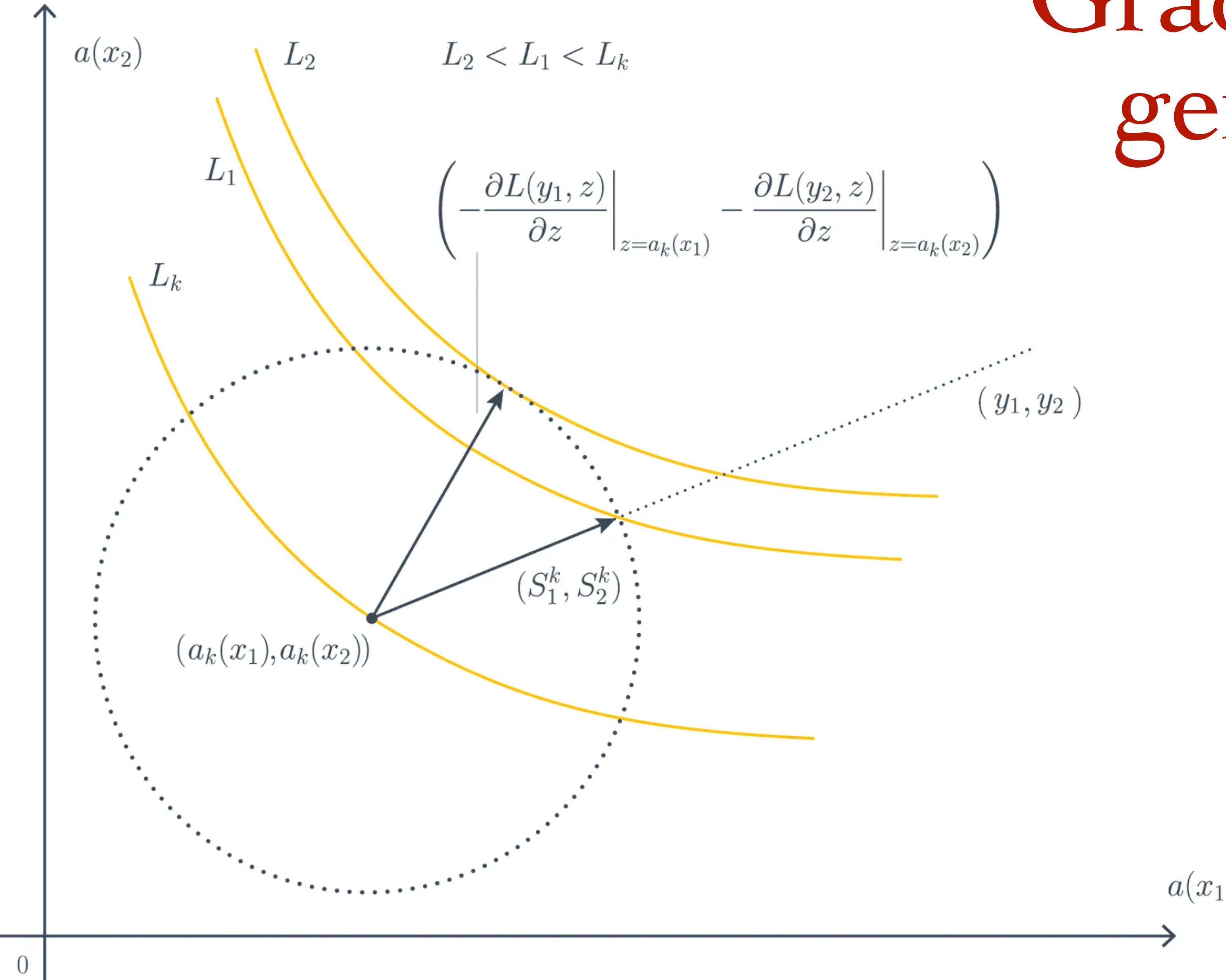
- Если  $y_i = a_{N-1}(x_i)$ , то объект не участвует в обучении
- Если  $y_i \neq a_{N-1}(x_i)$ , то базовая модель учится выдавать корректный класс

# GB / логистическая регрессия

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log \left( 1 + \exp \left( -\frac{y_i - a_{N-1}(x_i)}{2} b_N(x_i) \right) \right) \rightarrow \min_{b_N(x)}$$

- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -0.5 \rightarrow$  надо  $b_N(x_i) > 0.5$
- $y_i = +1, \sum_{n=1}^{N-1} b_n(x_i) = -100 \rightarrow$  надо  $b_N(x_i) > 100$
- Но на обоих объектах будет одинаково максимизироваться отступ
- На объектах с корректными ответами никак не контролируется выход  $b_N(x)$

# Gradient Boosting: general scheme



# **GB/ почему градиентный?**

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow \min_{b_N(x)}$$

- Ошибка на объекте  $x_i$  при прогнозе новой модели, равном  $z$  :

$$L(y_i, a_{N-1}(x_i) + z)$$

- Как посчитать, куда и как сильно сдвигать  $a_{N-1}(x_i)$ , чтобы уменьшить ошибку?

# GB/ почему градиентный?

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow \min_{b_N(x)}$$

- Ошибка на объекте  $x_i$  при прогнозе новой модели, равном  $z$  :

$$L(y_i, a_{N-1}(x_i) + z)$$

- Как посчитать, куда и как сильно сдвигать  $a_{N-1}(x_i)$ , чтобы уменьшить ошибку?
- Посчитать производную

# **GB/ почему градиентный?**

- Ошибка на объекте  $x_i$  при прогнозе новой модели, равном  $z$  :

$$L(y_i, a_{N-1}(x_i) + z)$$

- Посчитаем производную:

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)}$$

# GB/ почему градиентный?

- Посчитаем производную:

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)}$$

- Знак показывает, в какую сторону сдвигать прогноз на  $x_i$ , чтобы уменьшить ошибку композиции на нём
- Величина показывает, как сильно можно уменьшить ошибку, если сдвинуть прогноз
- Если ошибка почти не сдвинется, то нет смысла что-то менять

# GB/ почему градиентный?

- Обучение  $N$ -й модели:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_N(x_i) - s_i^{(N)} \right)^2 \rightarrow \min_{b_N(x)}$$

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)} \text{ — сдвиги}$$

- Как бы градиентный спуск в пространстве ответов на обучающей выборке
- Базовая модель будет делать корректировки на объектах так, чтобы как можно сильнее уменьшить ошибку композиции
- Сдвиги учитывают особенности функции потерь

# GB / шаг назад к MSE

$$\frac{1}{\ell} \sum_i L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow \min_{b_N}$$

$$L(y, z) = \frac{1}{2}([z > y_i] \cdot 10 + [z < y_i] \cdot 1) \cdot (z - y_i)^2$$

$$\frac{1}{\ell} \sum_i (b_N(x_i) - 10(y_i - a_{N-1}(x_i)))^2 \rightarrow \min_{b_N}$$

$$\begin{aligned} s_i^{(N)} &= -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)} = -\frac{\partial}{\partial z} \frac{1}{2} (z - y_i)^2 \Big|_{z=a_{N-1}(x_i)} = \\ &= -(a_{N-1}(x_i) - y_i) = y_i - a_{N-1}(x_i) \end{aligned}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (b_N(x_i) - (y_i - a_{N-1}(x_i)))^2 \rightarrow \min_{b_N(x)}$$

# GB/ а что с логистической регрессией?

$$\begin{aligned}s_i^{(N)} &= -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)} = \\&= -\frac{\partial}{\partial z} \log(1 + \exp(-y_i z)) \Big|_{z=a_{N-1}(x_i)} = \\&= \frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))}\end{aligned}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_N(x_i) - \frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \right)^2 \rightarrow \min_{b_N(x)}$$

# GB/ а что с логистической регрессией?

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \left( b_N(x_i) - \frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \right)^2 \rightarrow \min_{b_N(x)}$$

- Отступ большой положительный:  $\frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \approx 0$
- Отступ большой отрицательный:  $\frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))} \approx \pm 1$
- Чтобы учесть особенности функции потерь, можно посчитать её производные в точке текущего прогноза композиции
- Базовую модель будем обучать на эти производные (со знаком минус)