МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 17

Дилемма «Смещение–разброс» и кривая обучения

Bias-variance trade-off и Learning curve

17.1. Дилемма «Смещение–разброс»

$$Y = f^*(x) + \varepsilon(x), \qquad \mathsf{E}\,\varepsilon(x) = 0, \qquad \mathsf{Var}\;\varepsilon(x) = \sigma_\varepsilon^2(x)$$

$f(x, D)$ – решающее правило,
$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$ – обучающая выборка
x, $f^*(x)$ случайными не являются

Случайными являются: $\varepsilon(x)$ и D, причем считаем их независимыми

Тогда

$$\text{MSE}(x) = \mathsf{E}_{D,\varepsilon}\big(f(x, D) - Y\big)^2 = \sigma_\varepsilon^2(x) + \underbrace{\mathsf{E}_D^2\big(f(x, D) - f^*(x)\big)}_{\text{Bias}^2 x} + \underbrace{\mathsf{E}_D\big(f(x, D) - \mathsf{E}_D f(x, D)\big)^2}_{\text{Variance}\, x}$$

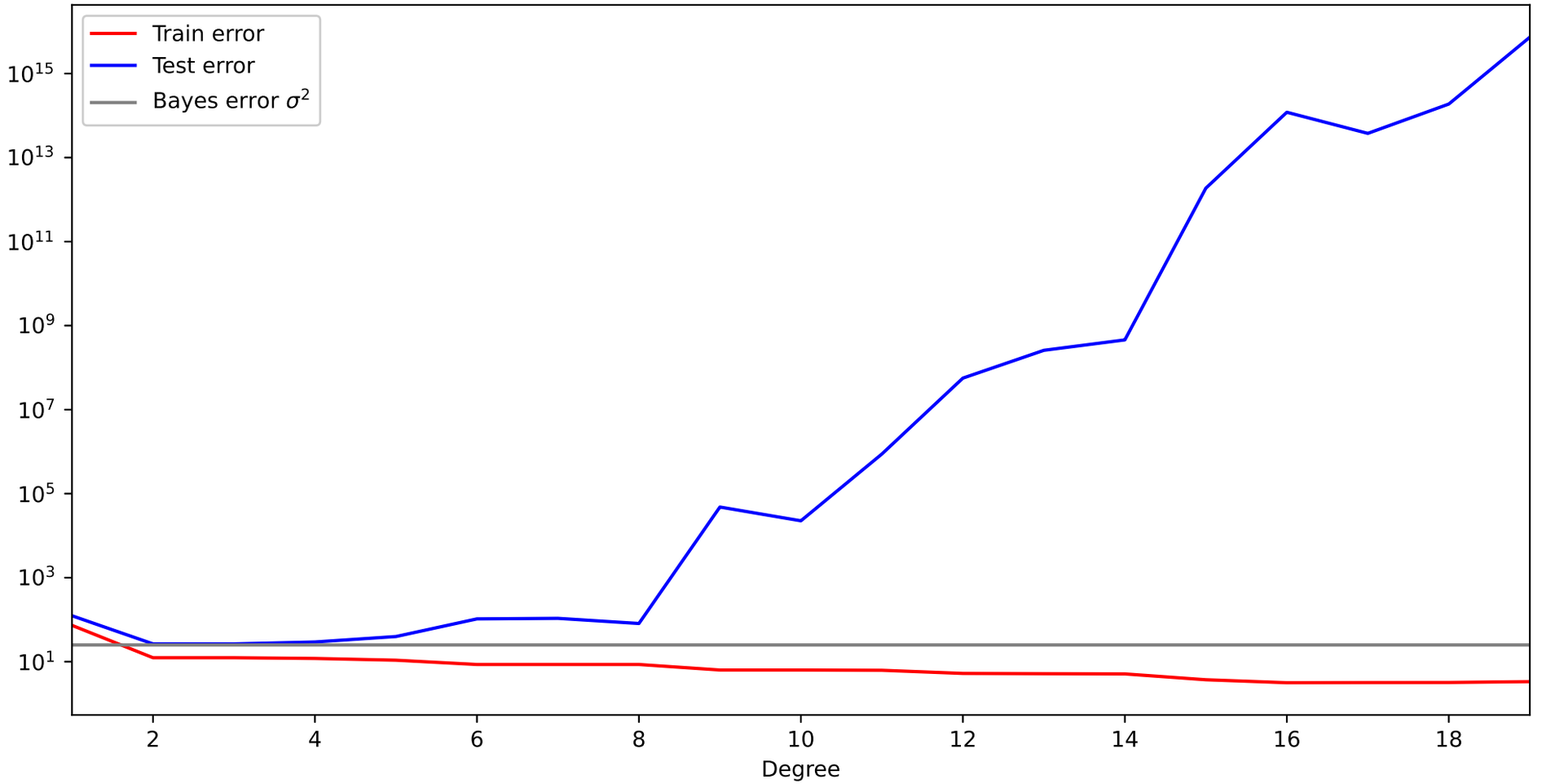
Ошибка = Неустранимая ошибка + Смещение^2 + Разброс

Доказательство:

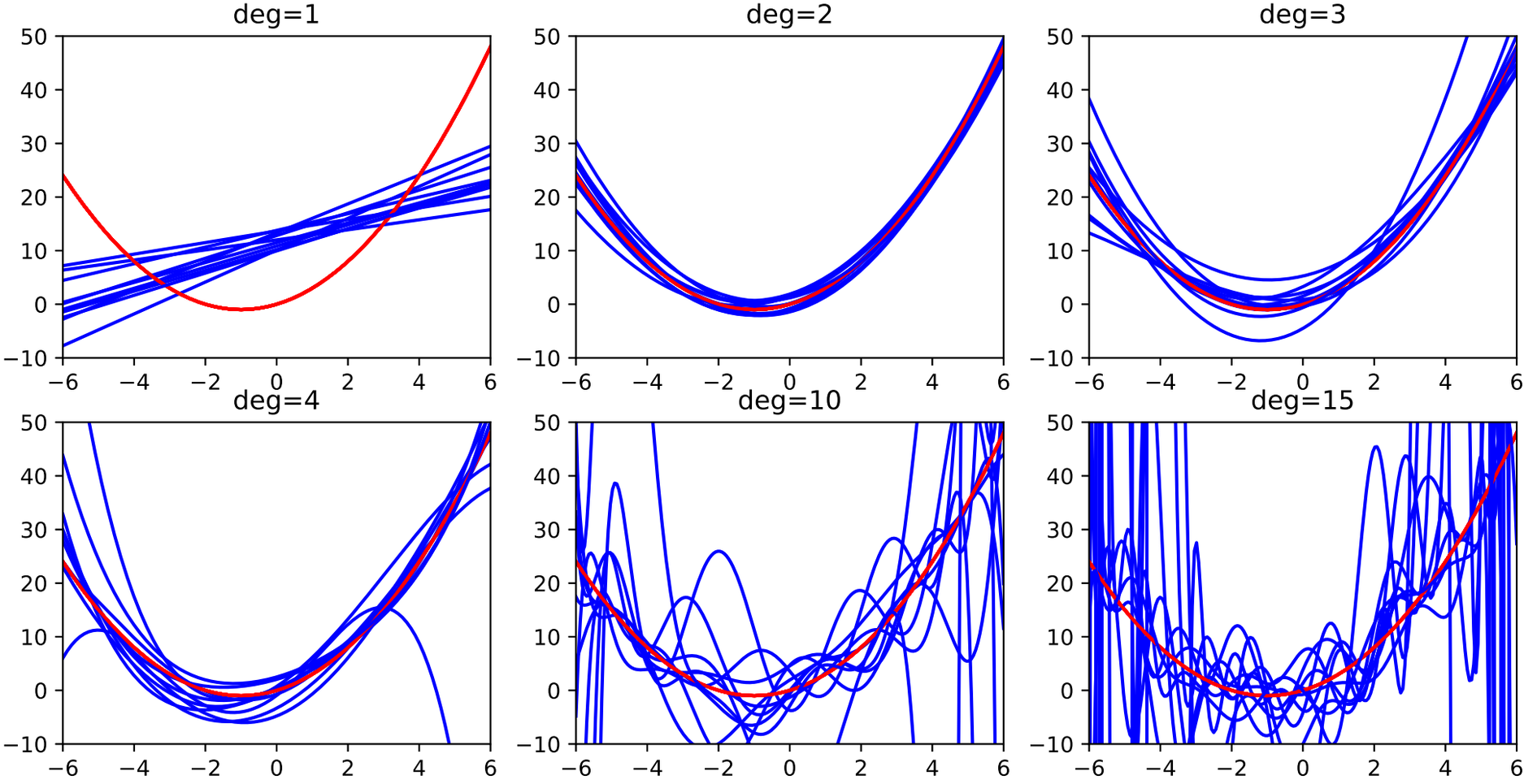
(можно раскрыть л.ч и п.ч. и увидеть, что они совпадают)

Большое смещение — недообучение. Большой разброс — переобучение.

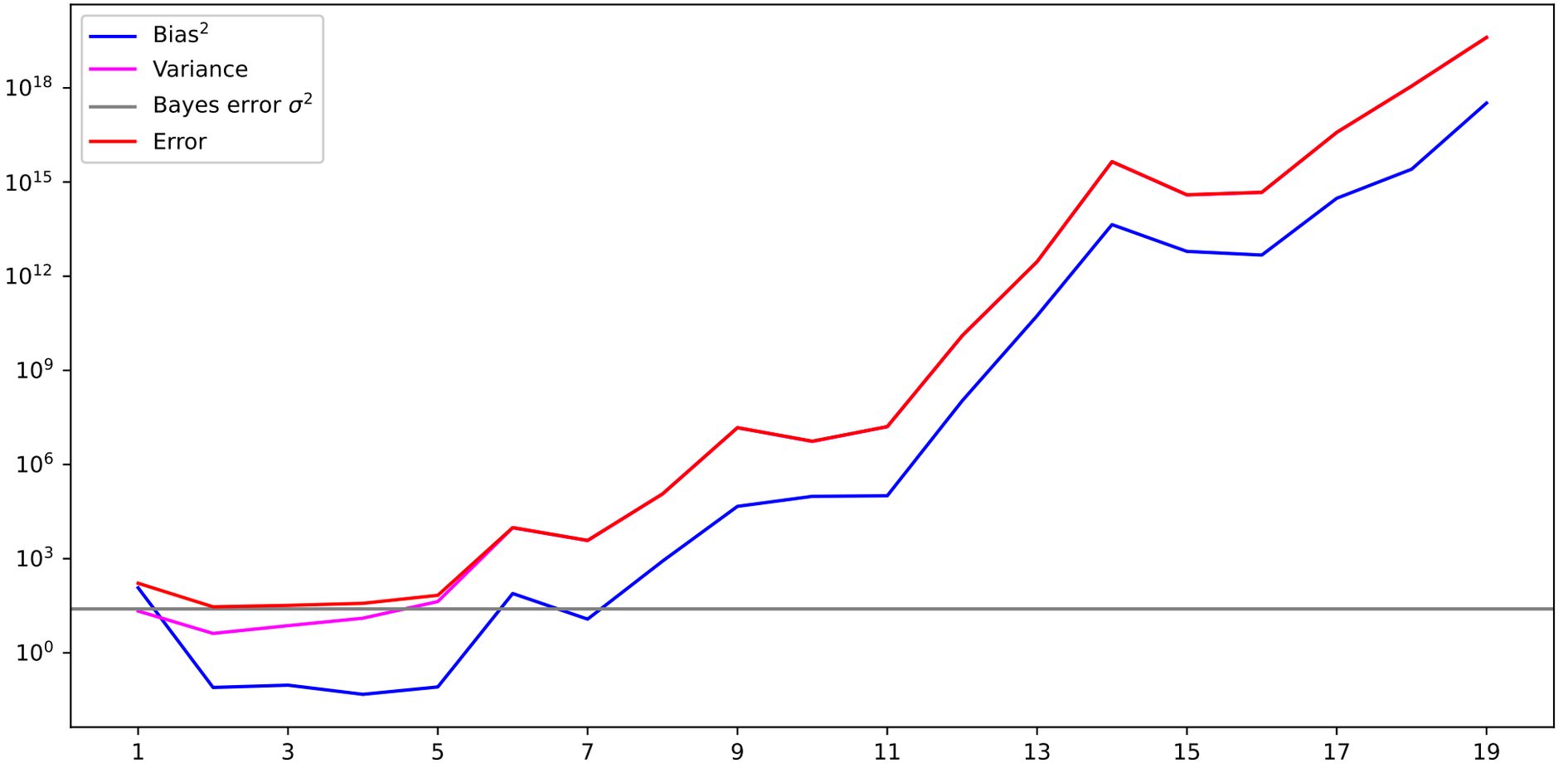
LSQ $y = x^2 + 2x + N(0, 5)$



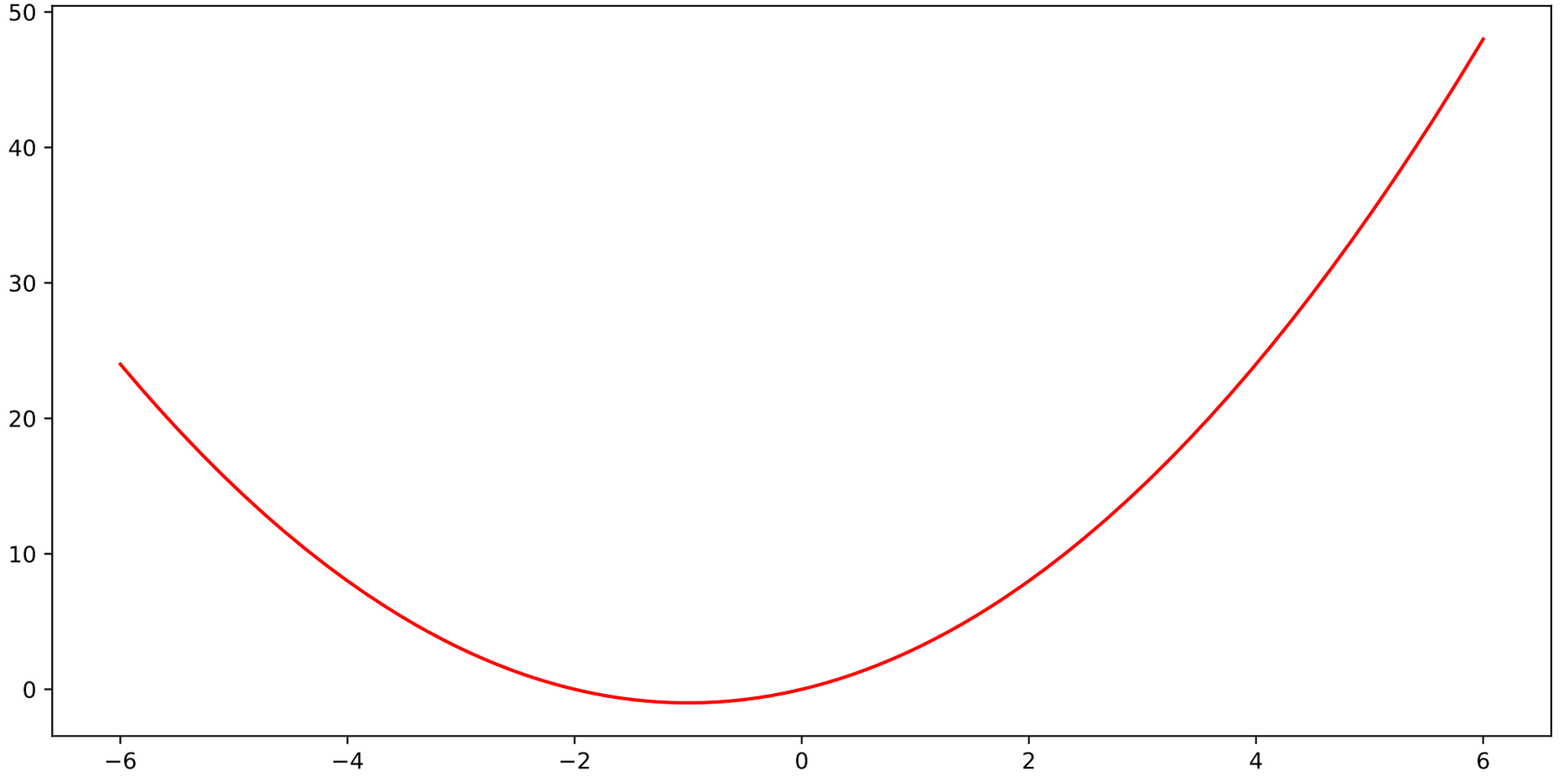
LSQ



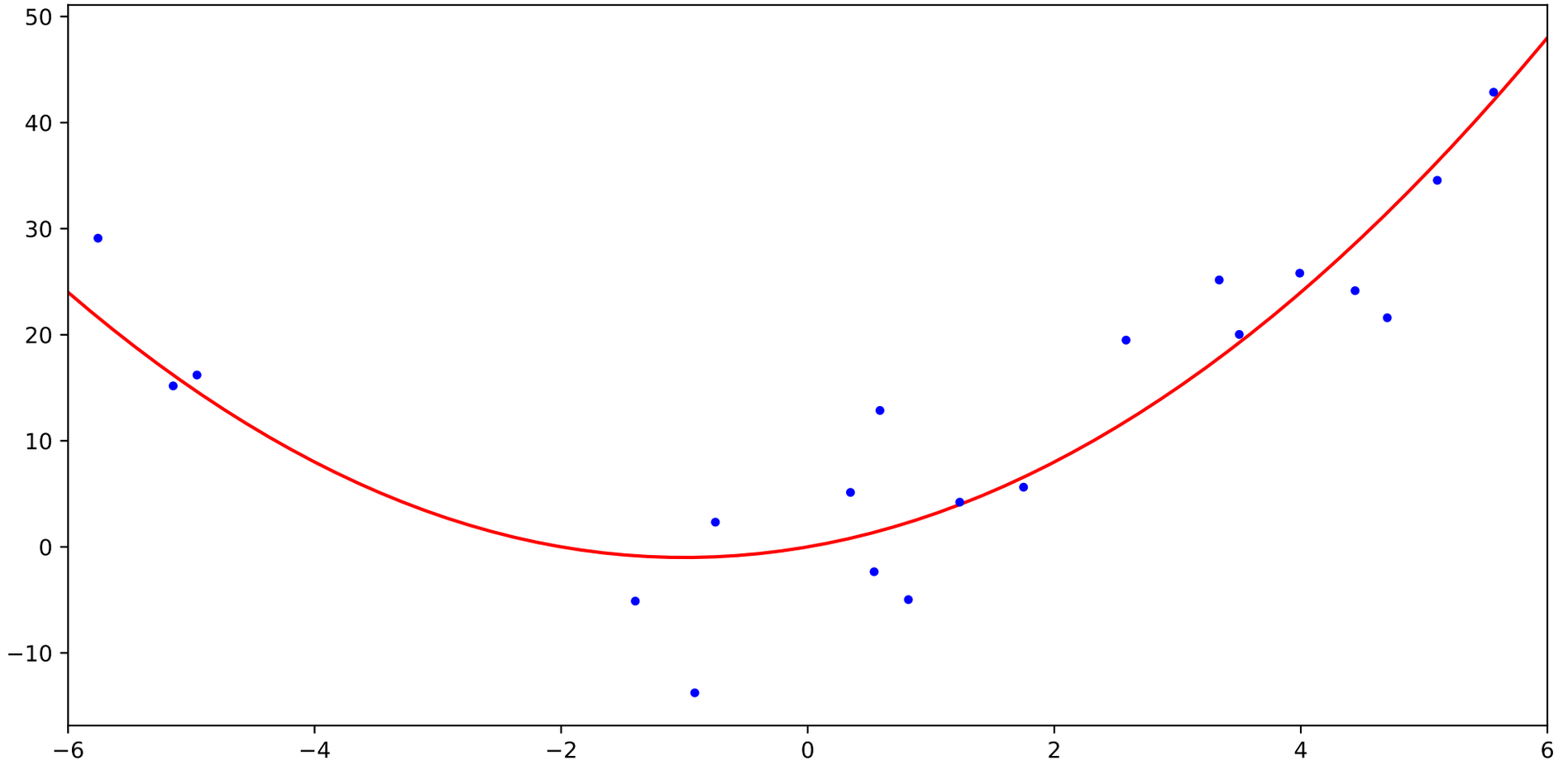
LSQ


Bias^2
Variance
Bayes error σ^2
Error
10^{18}
10^{15}
10^{12}
10^{9}
10^{6}
10^{3}
10^{0}
1
3
5
7
9
11
13
15
17
19

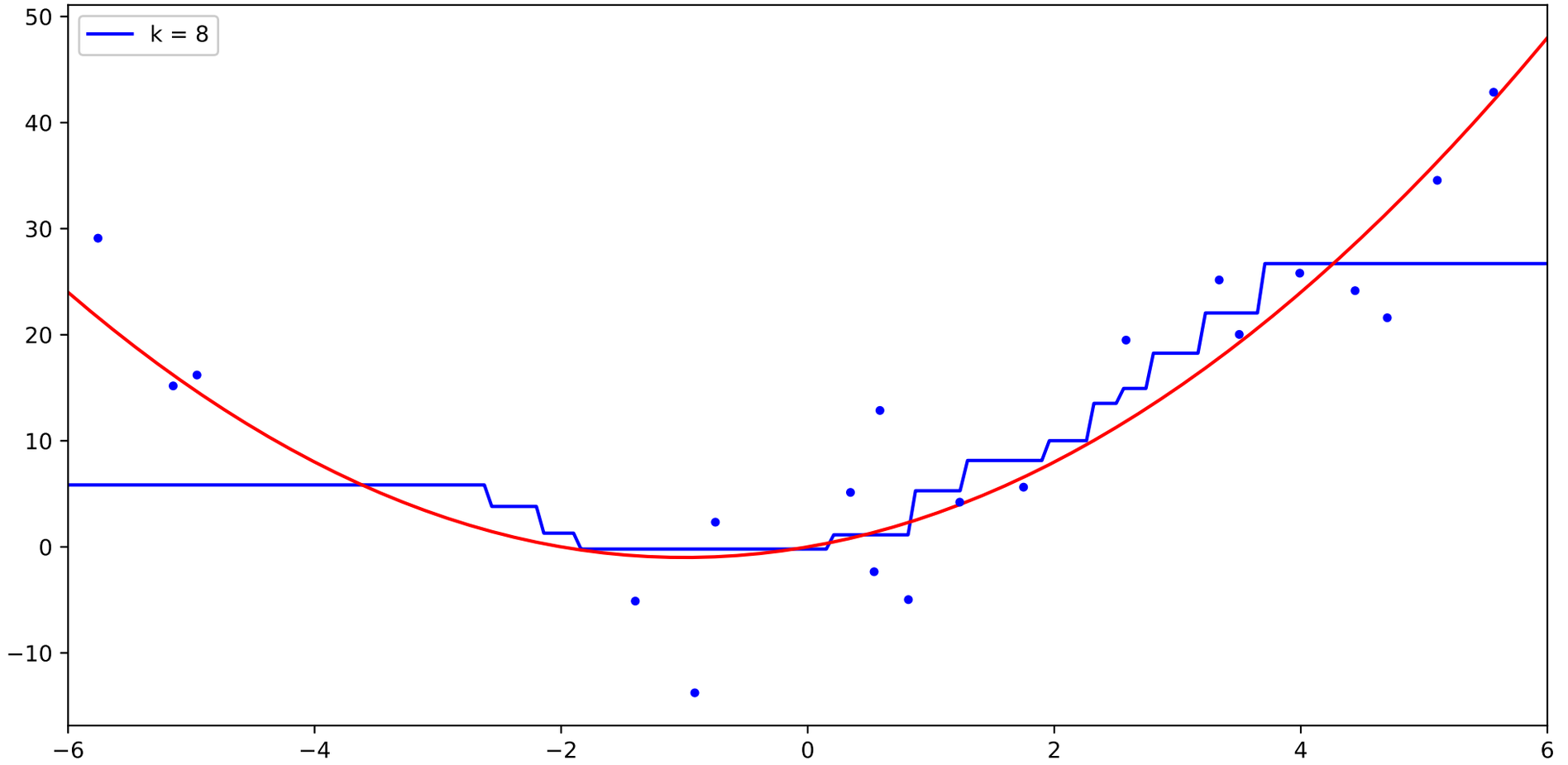
kNN $y = x^2 + 2x + N(0, 5)$



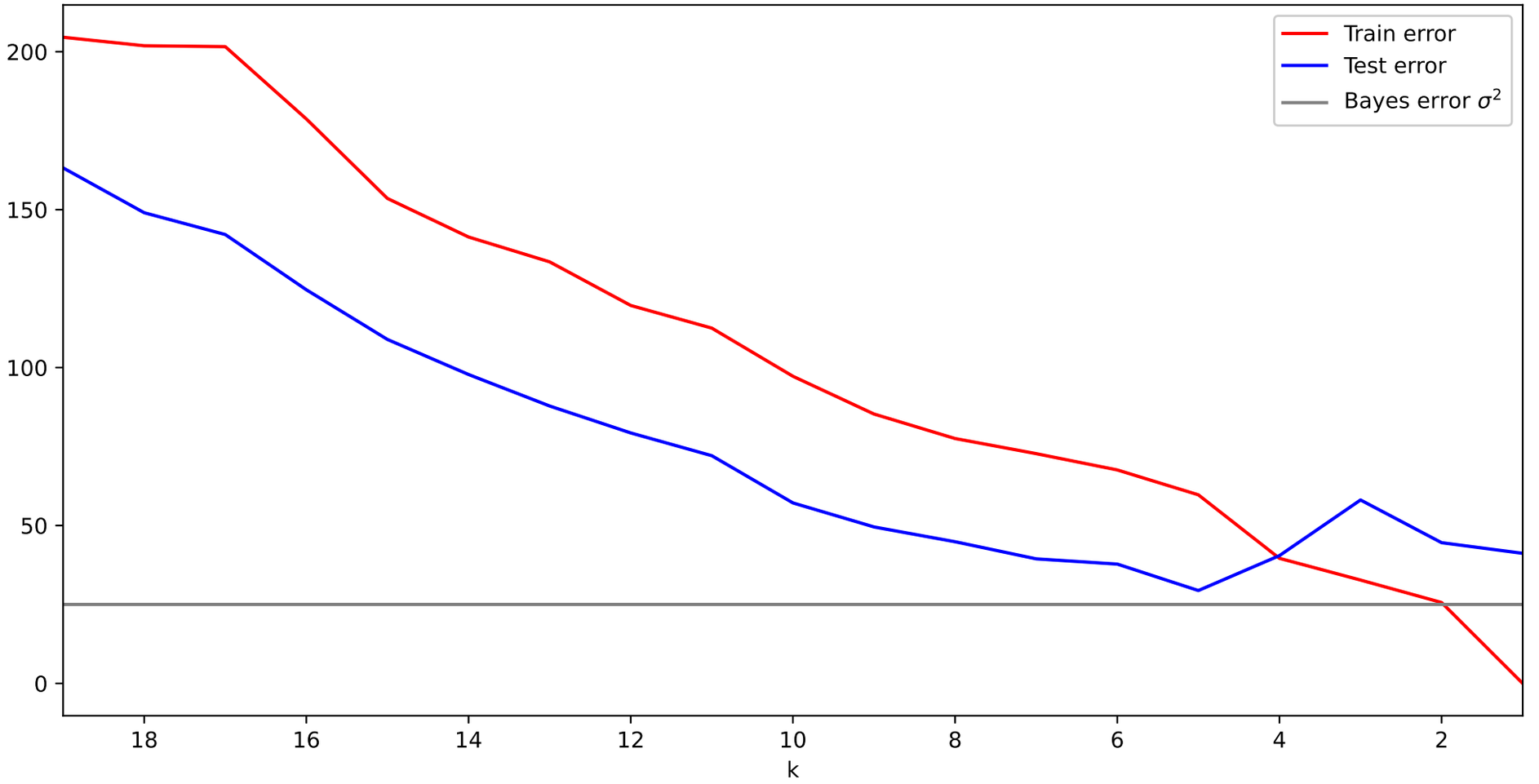
kNN $y = x^2 + 2x + N(0, 5)$

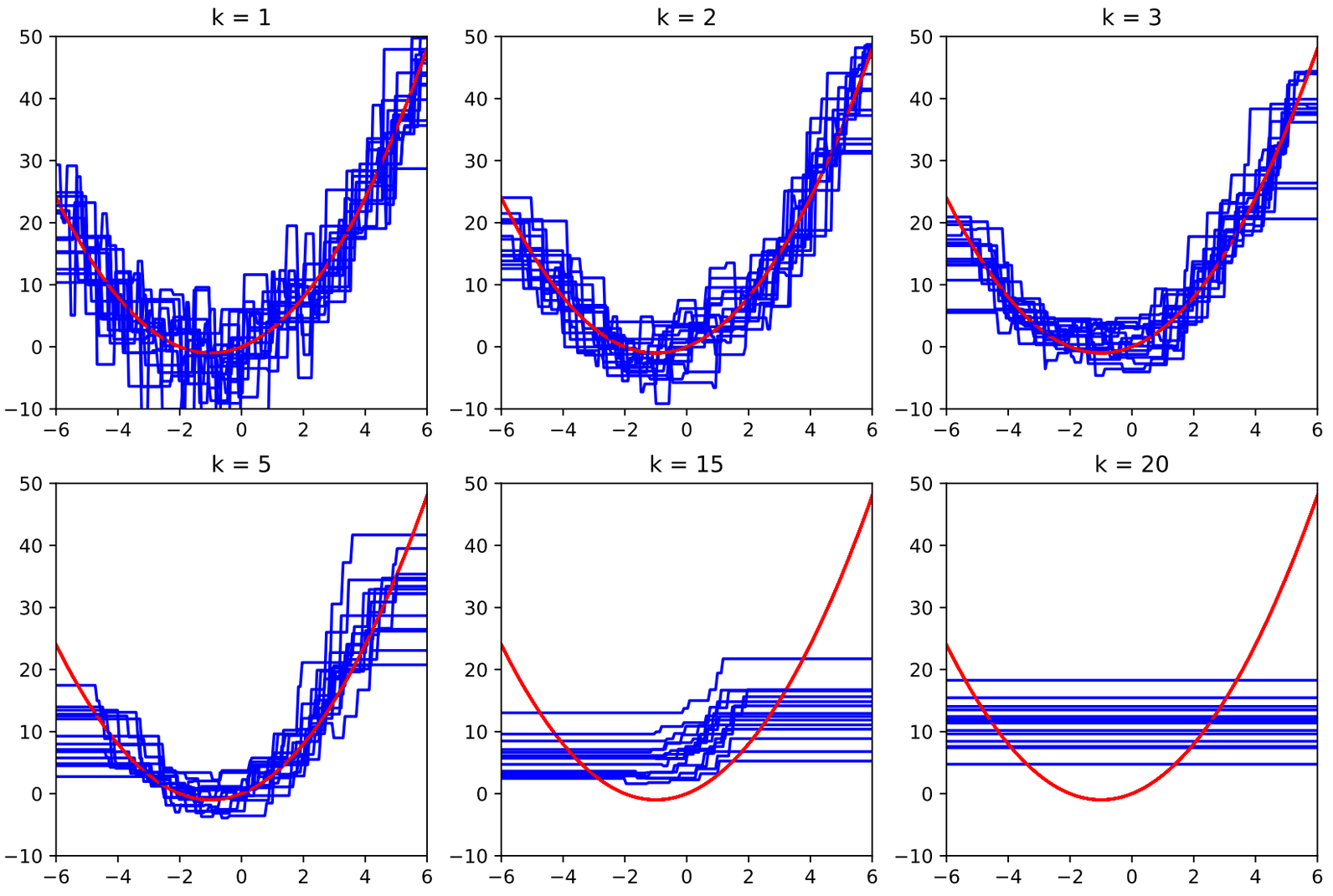


kNN $y = x^2 + 2x + N(0, 5)$

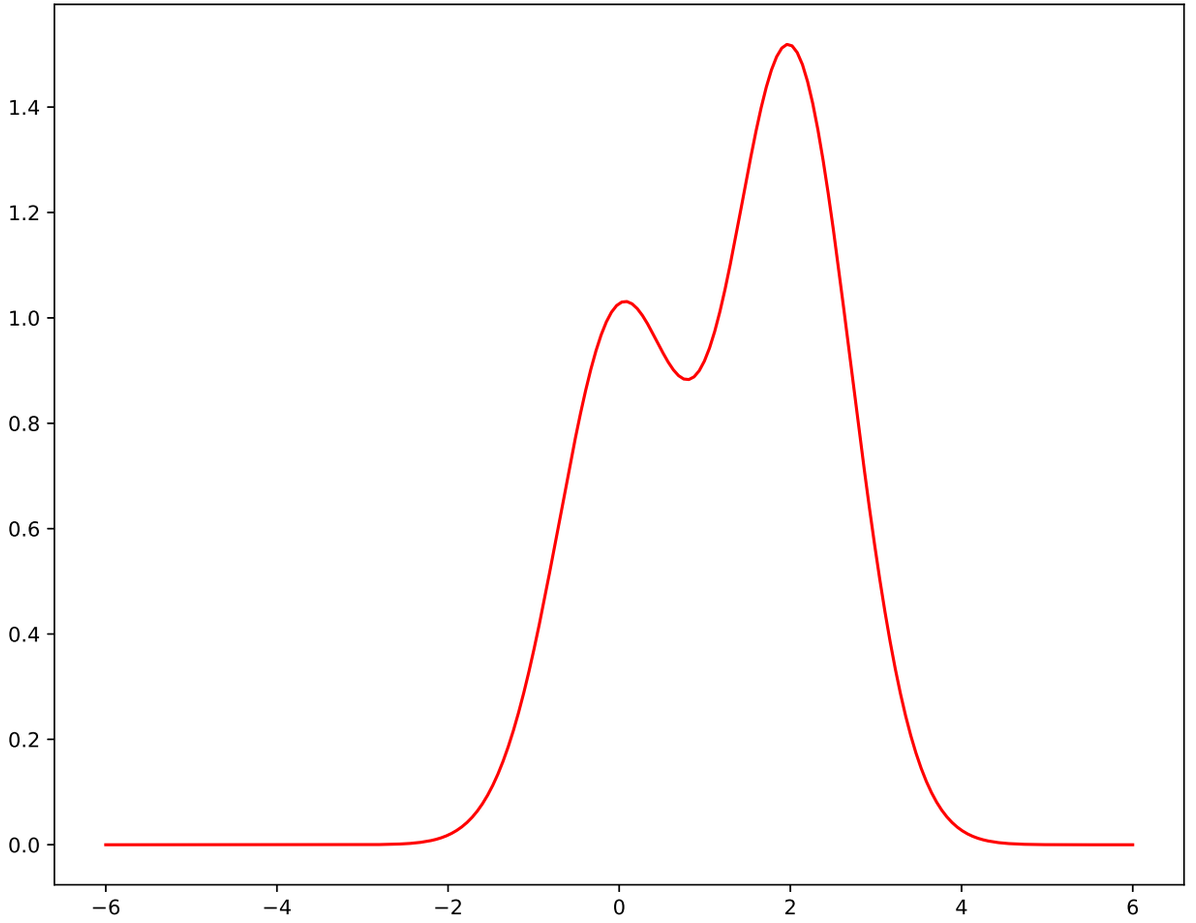


kNN $y = x^2 + 2x + N(0, 5)$

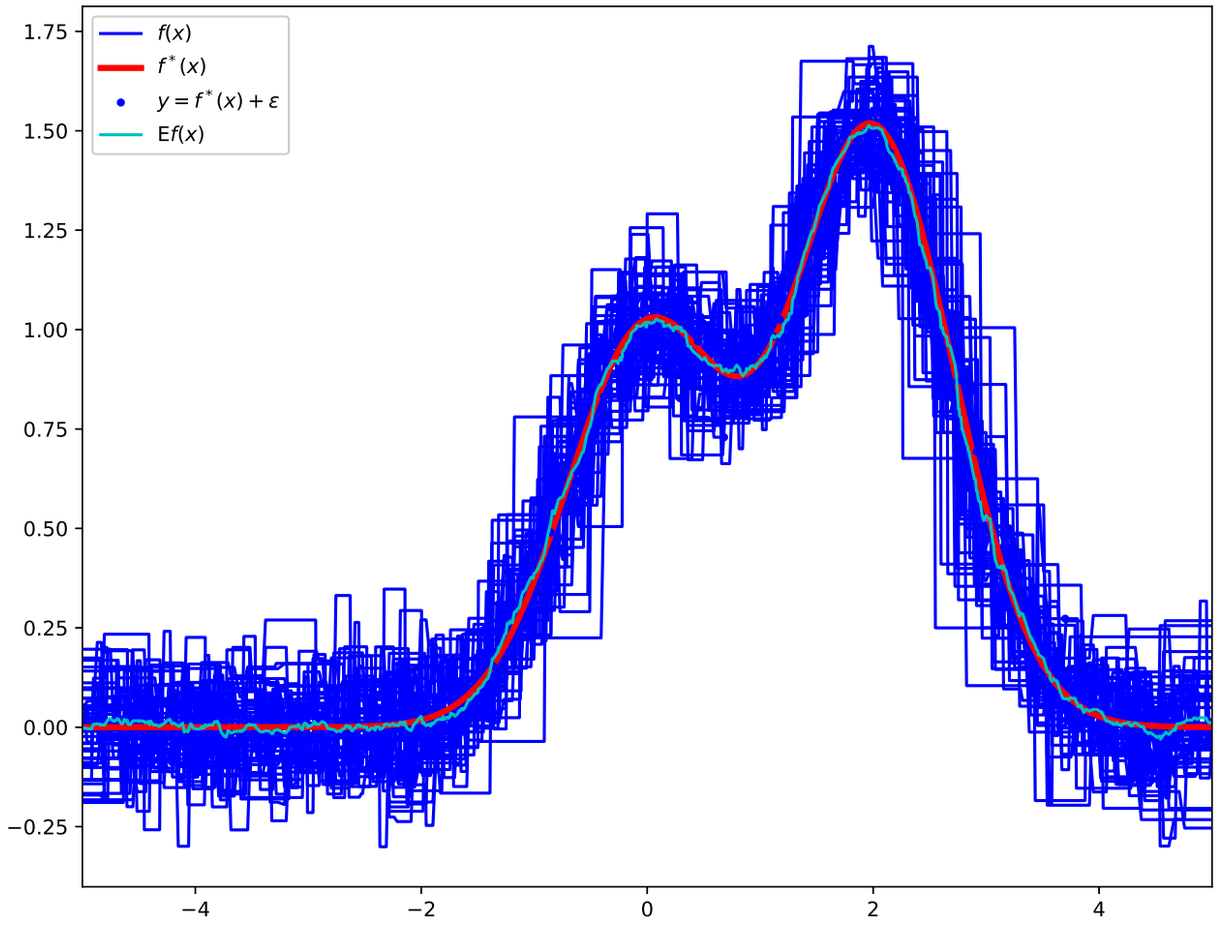



k = 1
k = 2
k = 3
k = 5
k = 15
k = 20
50
40
30
20
10
0
−10
−6
−4
−2
0
2
4
6

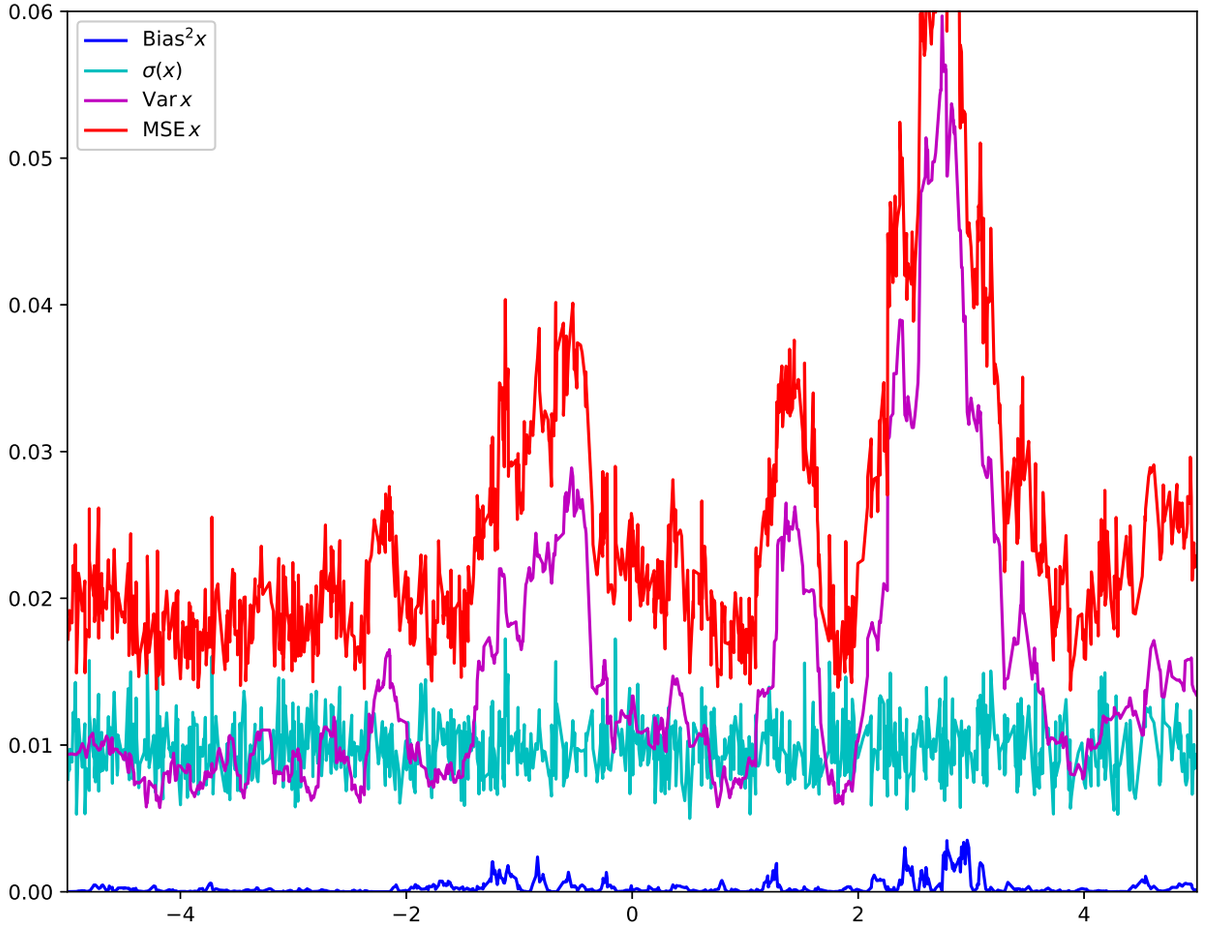
$$y = e^{-x^2} + 1.5e^{-(x-2)^2}$$



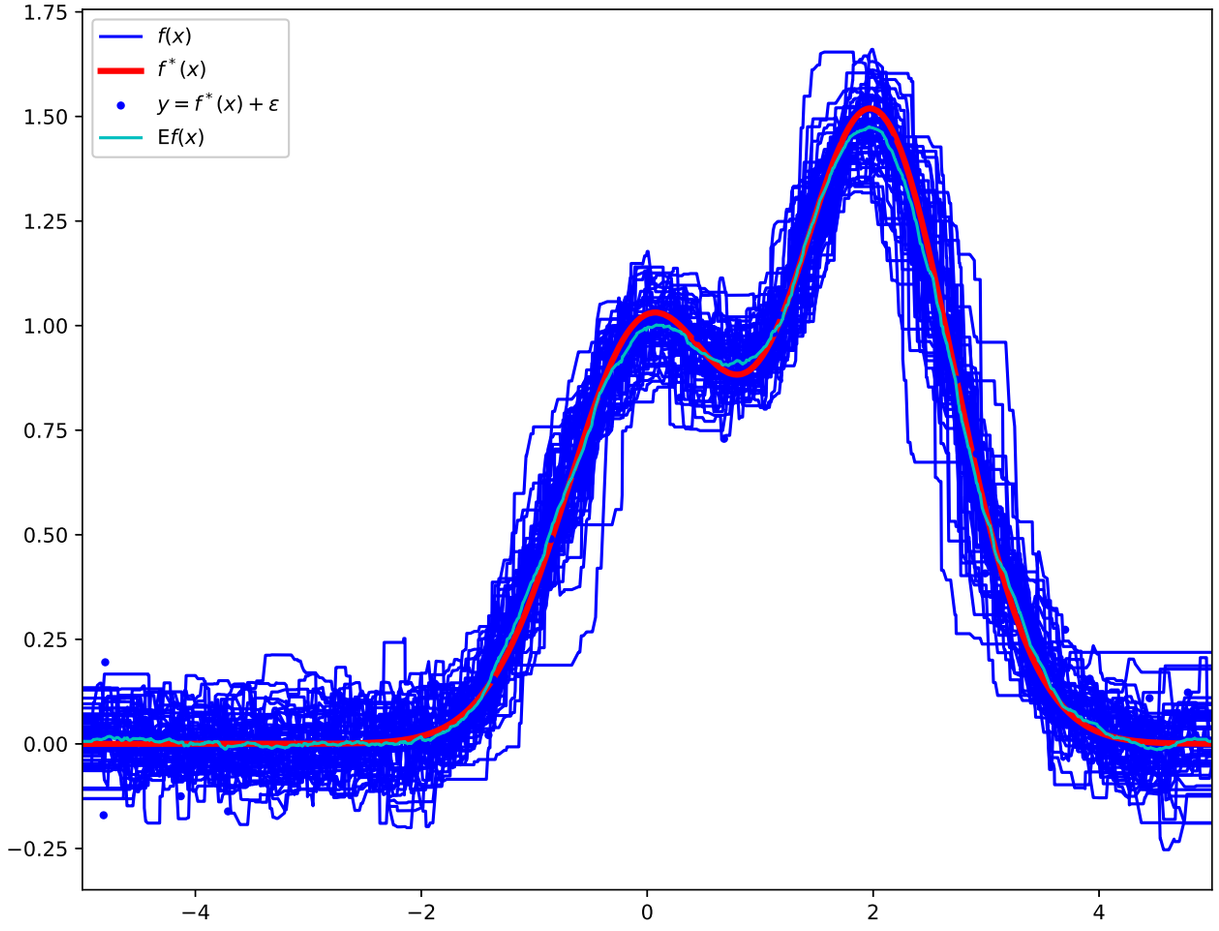
DT $y = e^{-x^2} + 1.5e^{-(x-2)^2}$



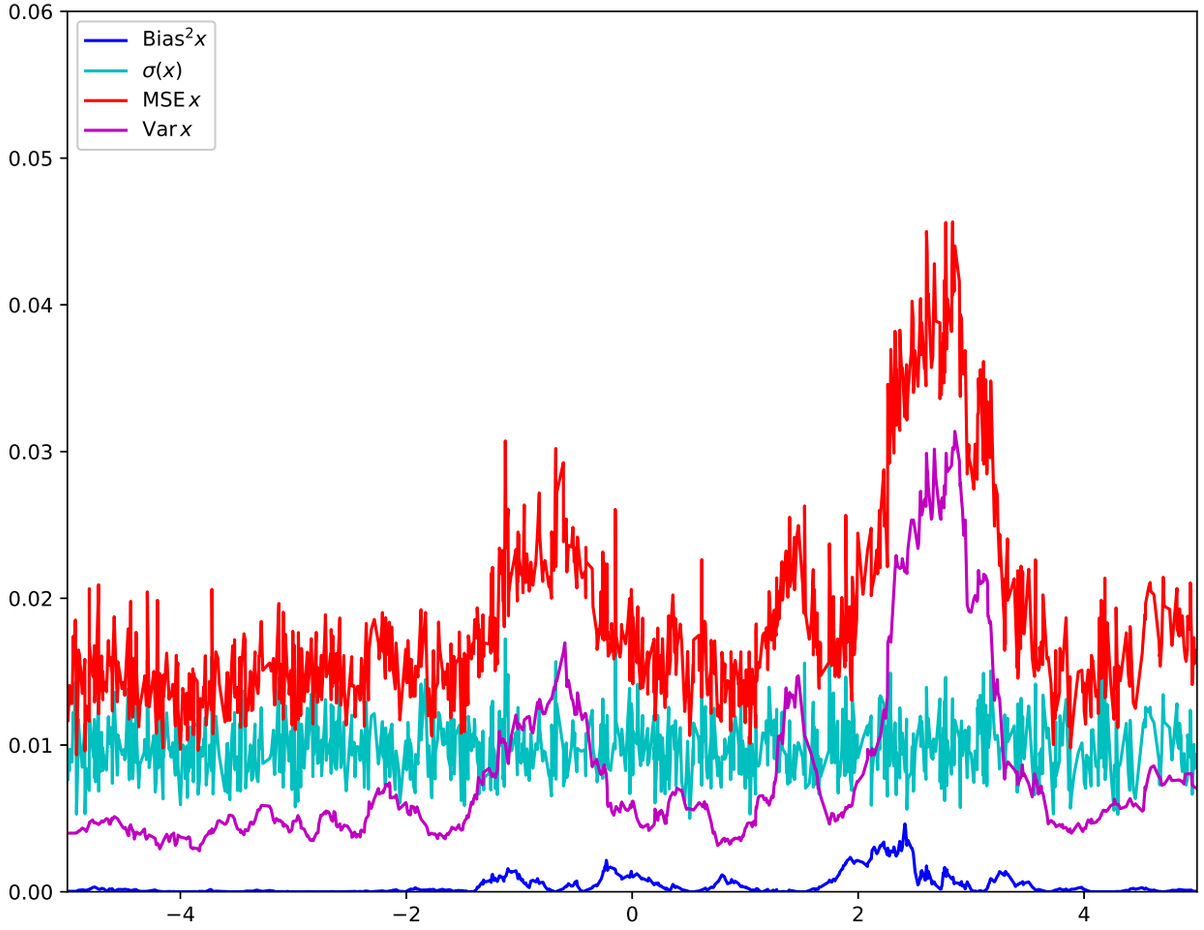
DT $y = e^{-x^2} + 1.5e^{-(x-2)^2}$



RF $y = e^{-x^2} + 1.5e^{-(x-2)^2}$



RF $y = e^{-x^2} + 1.5e^{-(x-2)^2}$



17.2. Кривая обучения (learning curve)