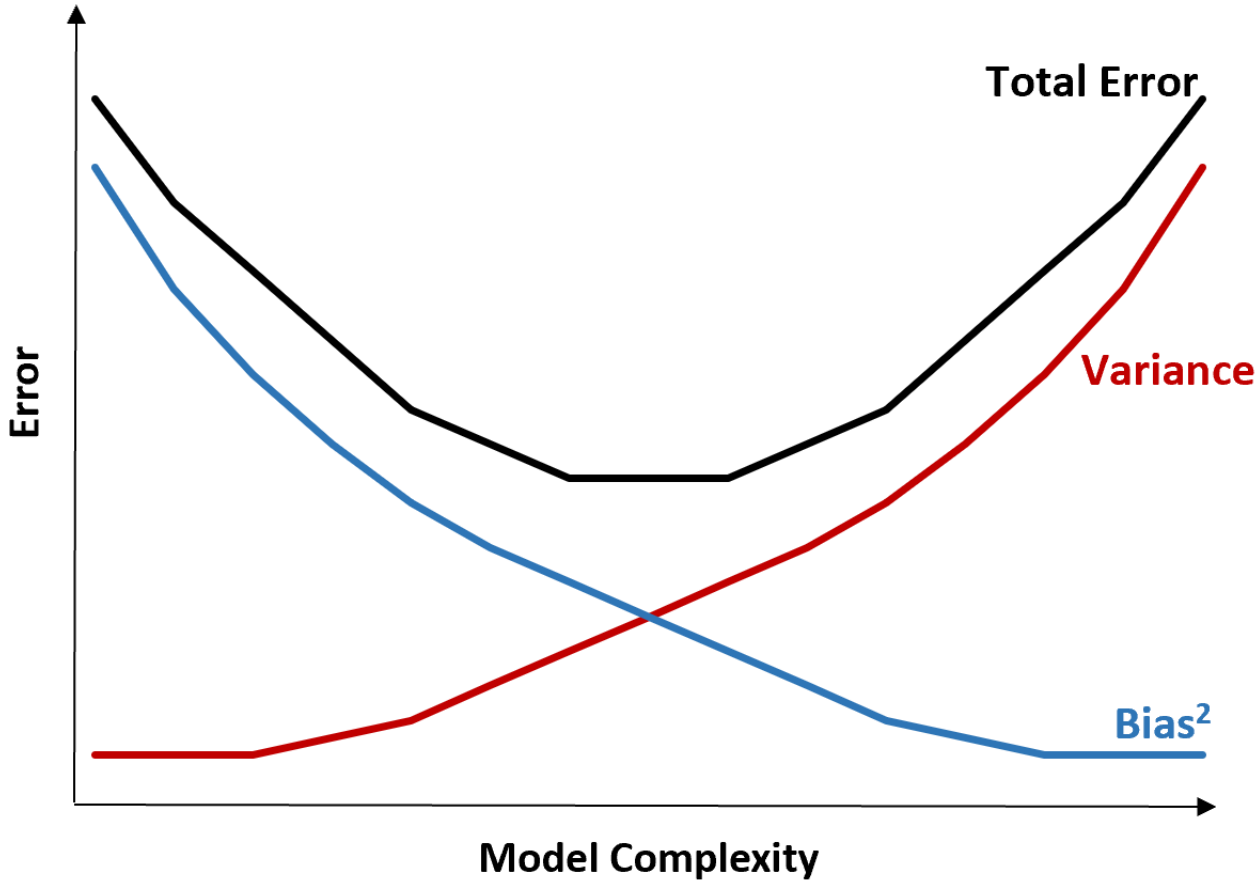
Double descent

Дмитриев Иван

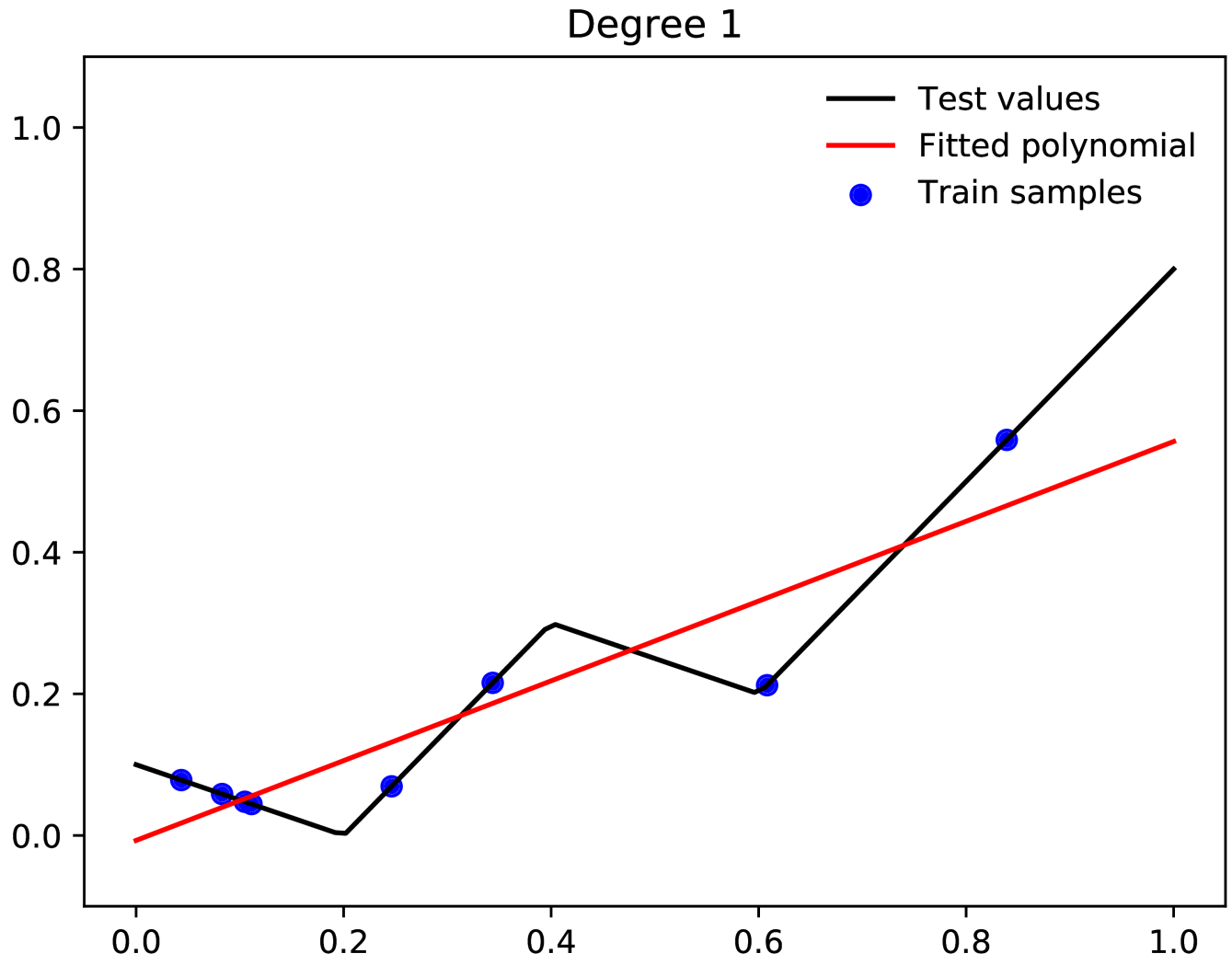
БПМИ202

Классический подход

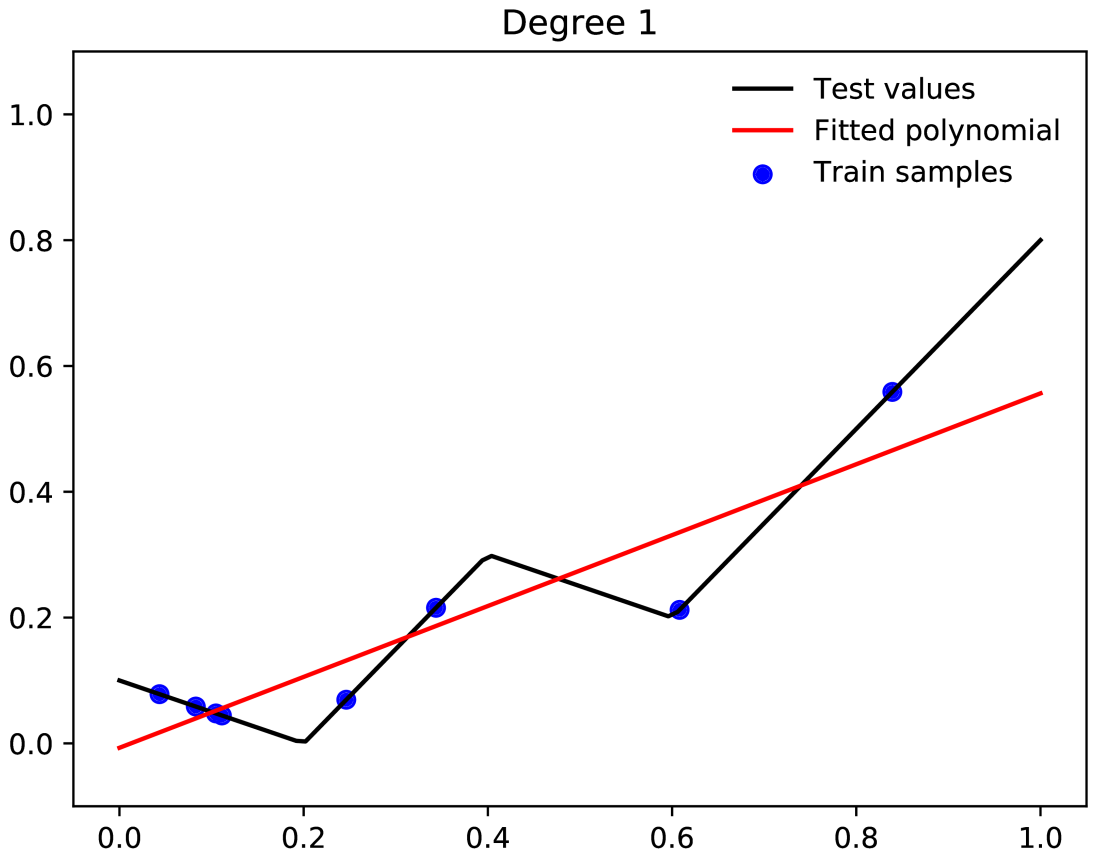


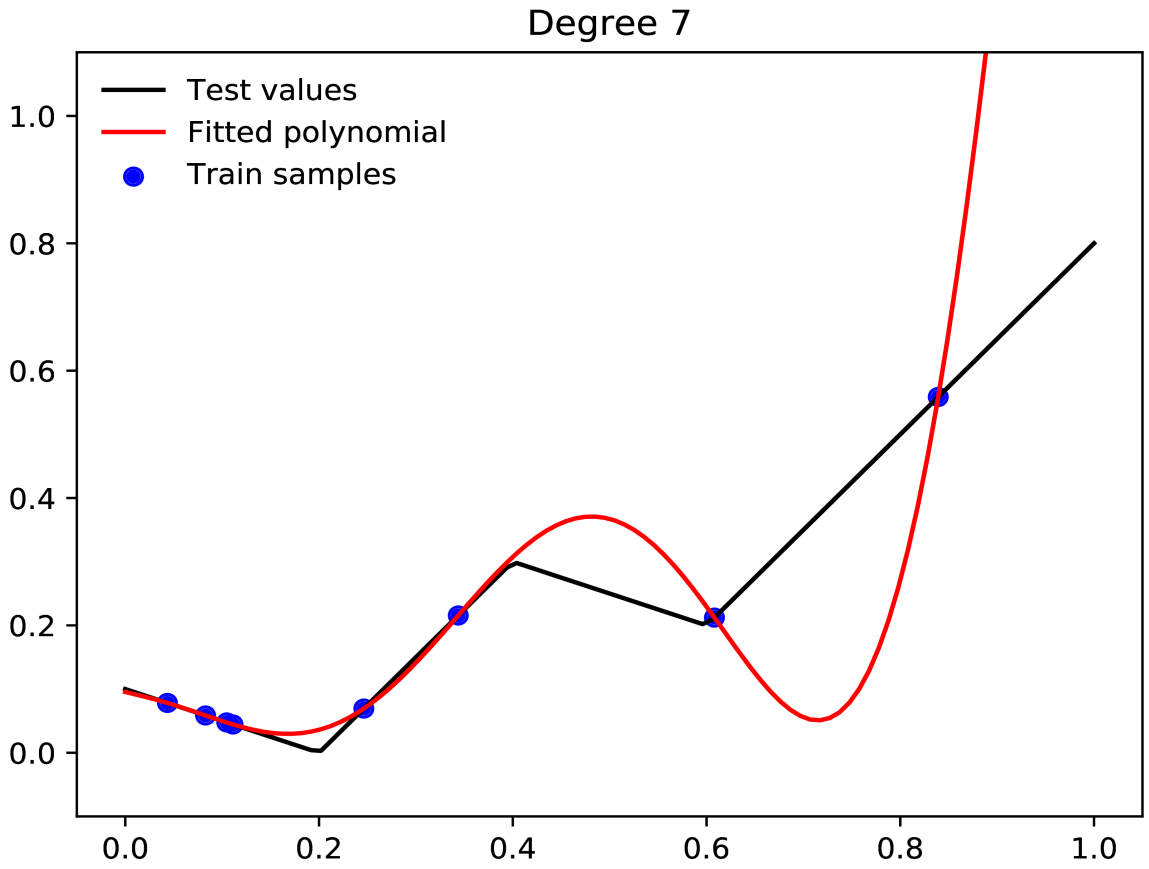
Пример

Возьмем несколько точек на кусочно-линейной функции
y =||x - 0.4| - 0.2| + x/2 - 0.1
Будем приближать результат многочленами

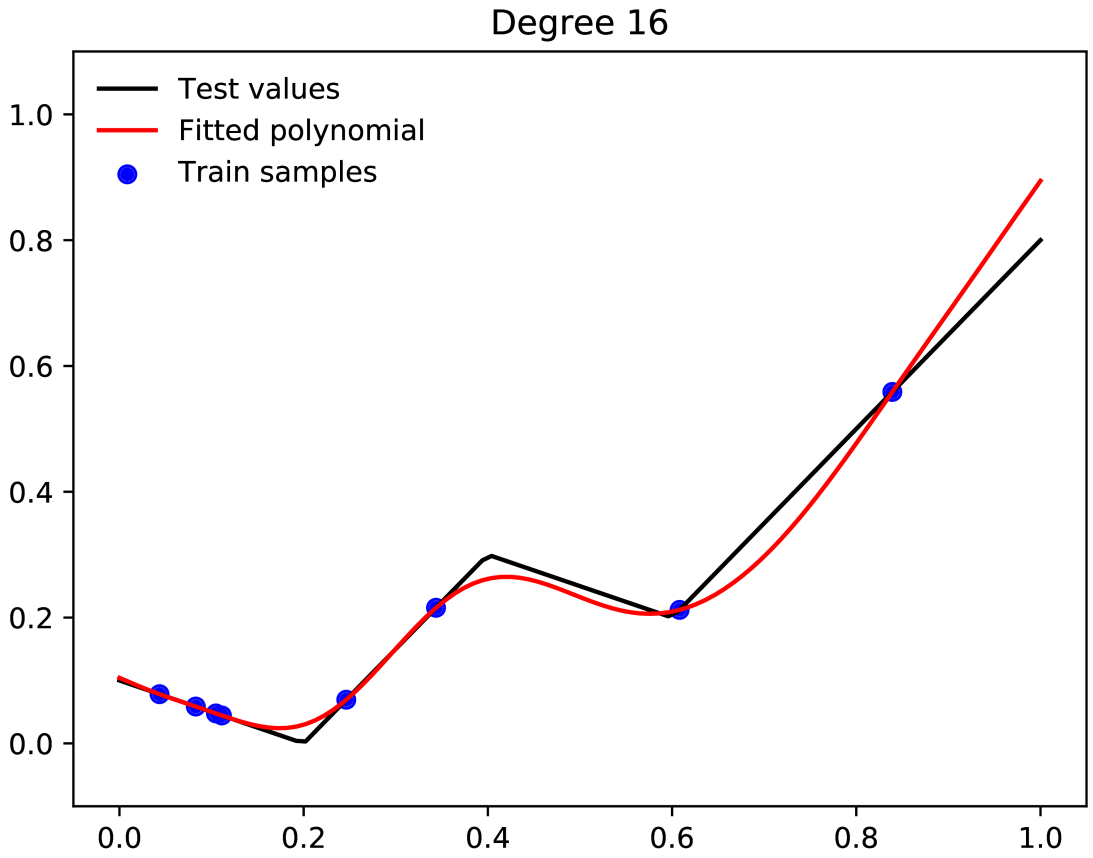


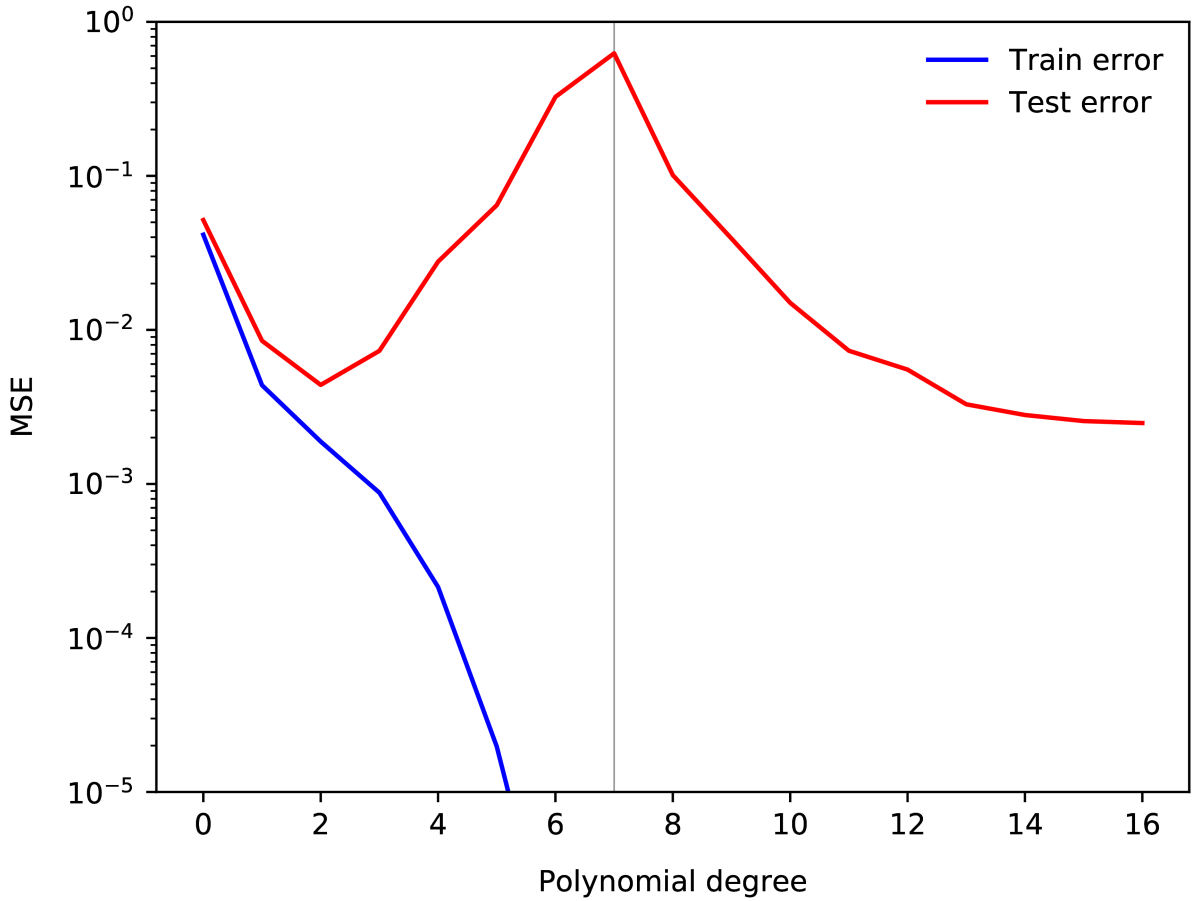
Пример


Degree 1
Test values
Fitted polynomial
Train samples
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

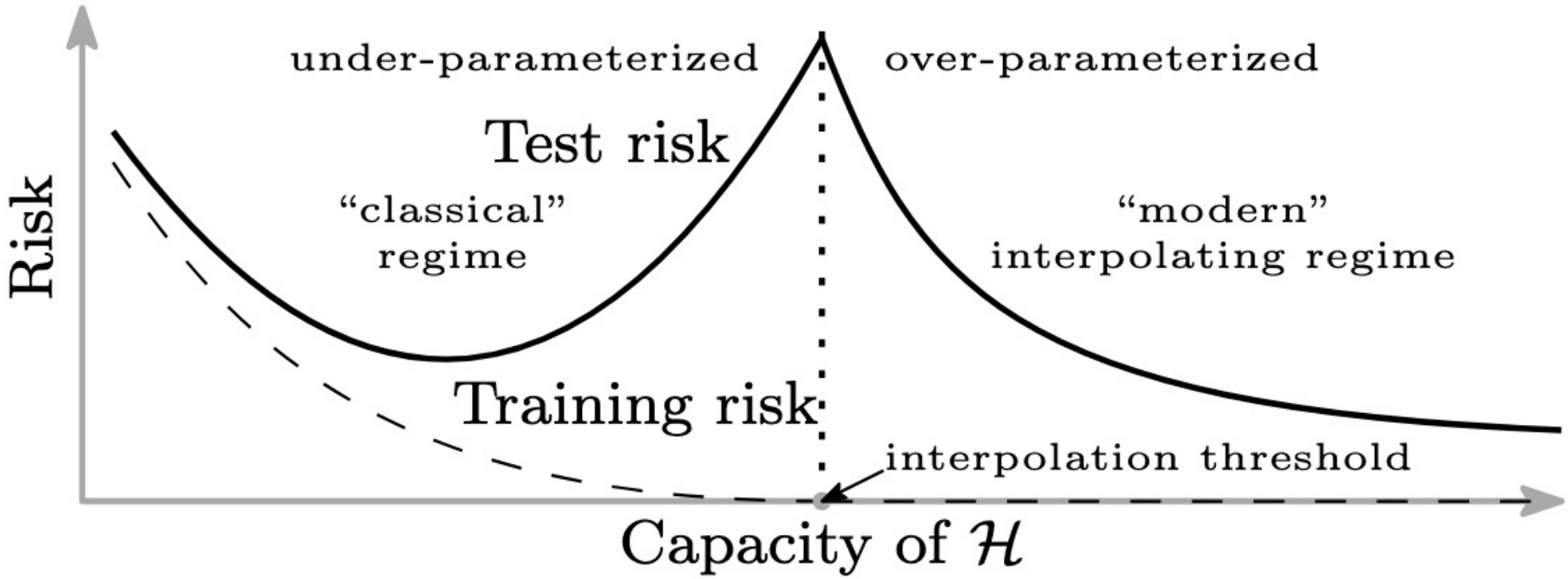

Degree 7
Test values
Fitted polynomial
Train samples
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Пример


Degree 16
Test values
Fitted polynomial
Train samples
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Train error
Test error
MSE
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
2
4
6
8
10
12
14
16
Polynomial degree

Double descent



Эффективная сложность модели

$$\mathrm{EMC}_{\mathrm{D},\varepsilon} = \max\{n \mid \mathbb{E}_{S\sim D^n}\left[Error_S\left(T(S)\right)\right] < \varepsilon\}$$

D – распределение на даннӹых
T – процедура обучения: принимает множество размеченных данных и возвращает классификатор
$\varepsilon > 0$

Гипотеза о двойном спуске

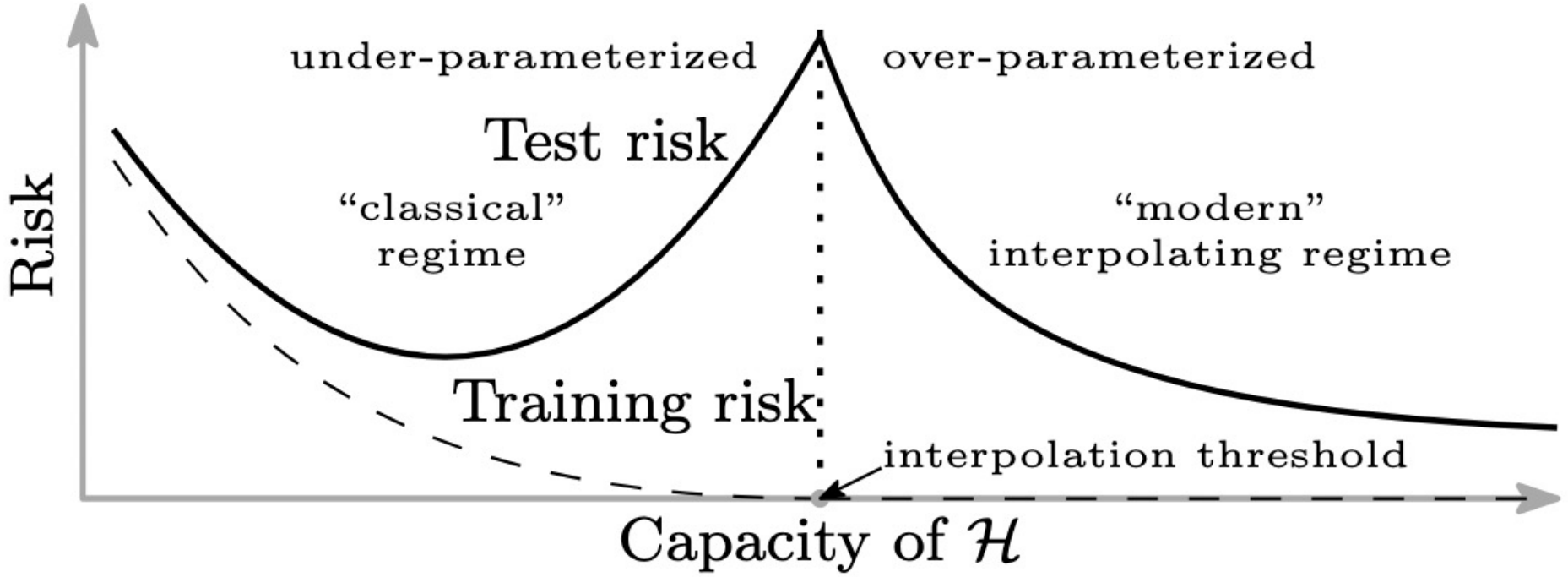
Пусть D – распределение на данных, T – процедура обучения, $\varepsilon > 0$, обучение происходит на n объектах. Тогда при увеличении эффективной модельной сложности T:

- Если $\mathrm{EMC}_{\mathrm{D},\varepsilon}$ < n, ошибка на тестовой выборке будет уменьшаться («классический» режим)
- Если $\mathrm{EMC}_{\mathrm{D},\varepsilon} \approx$ n, ошибка на тестовой выборке может и увеличиться, и уменьшиться
- Если $\mathrm{EMC}_{\mathrm{D},\varepsilon}$ > n, ошибка на тестовой выборке будет уменьшаться («современный» режим)

Гипотеза о двойном спуске: проблемы

- Непонятно, как подбирать ε, эвристически неплохо работает $\varepsilon = 0.1$
- Ширину «непонятного» диапазона непонятно, как определять

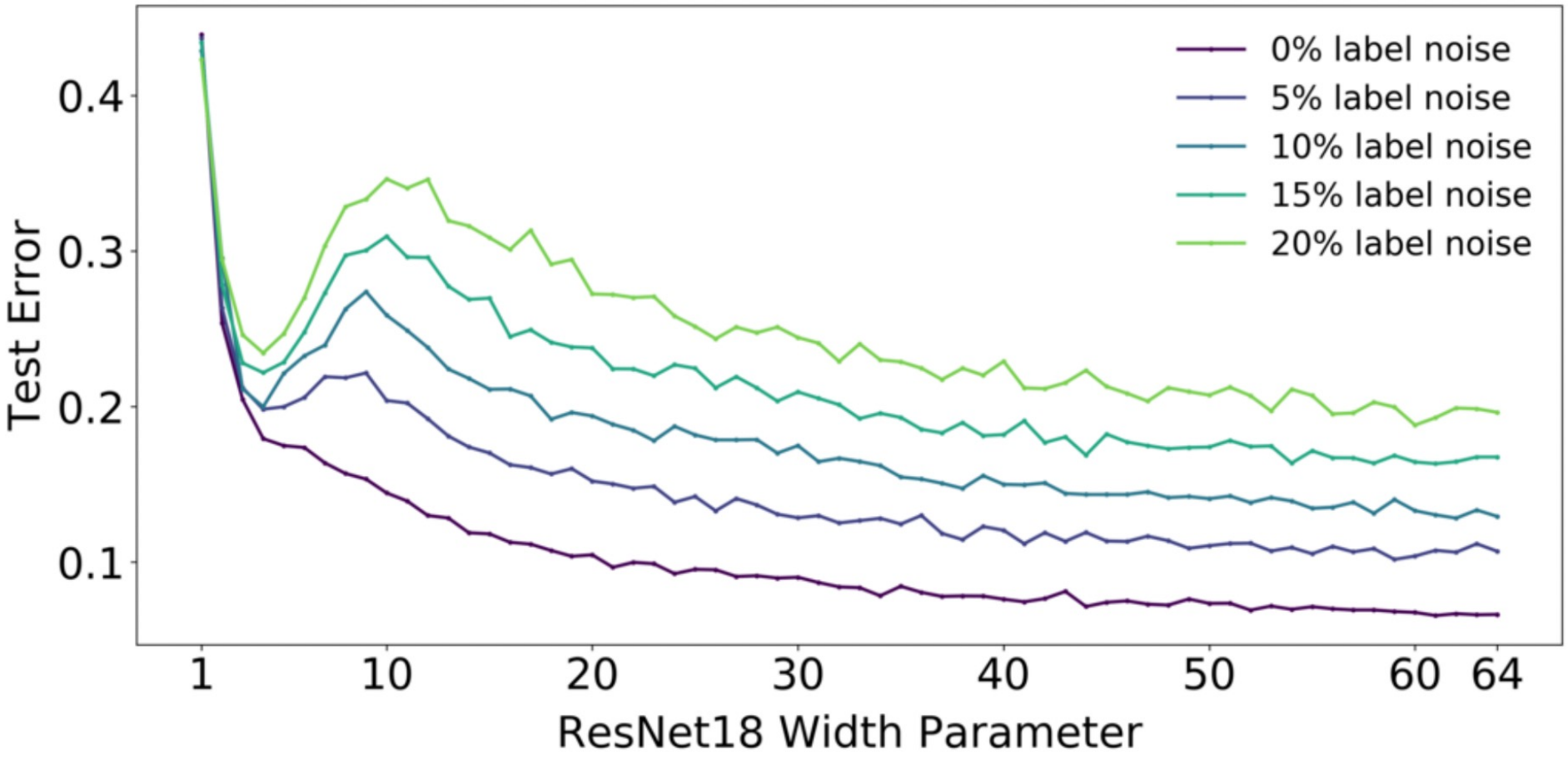
Гипотеза о двойном спуске



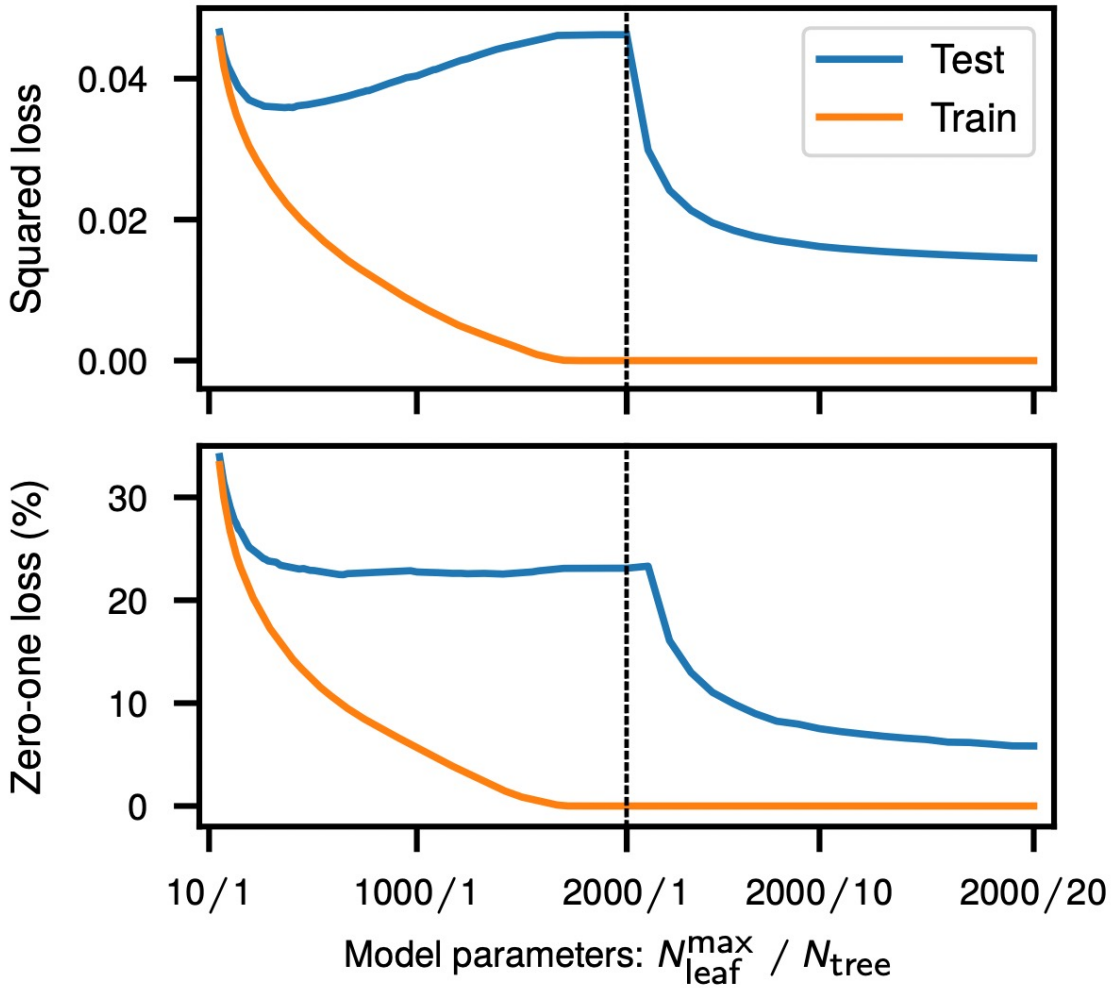
Где это появляется?

- Эффект наблюдается в совершенно разных сценариях
- Но практически исключительно на зашумленных данных

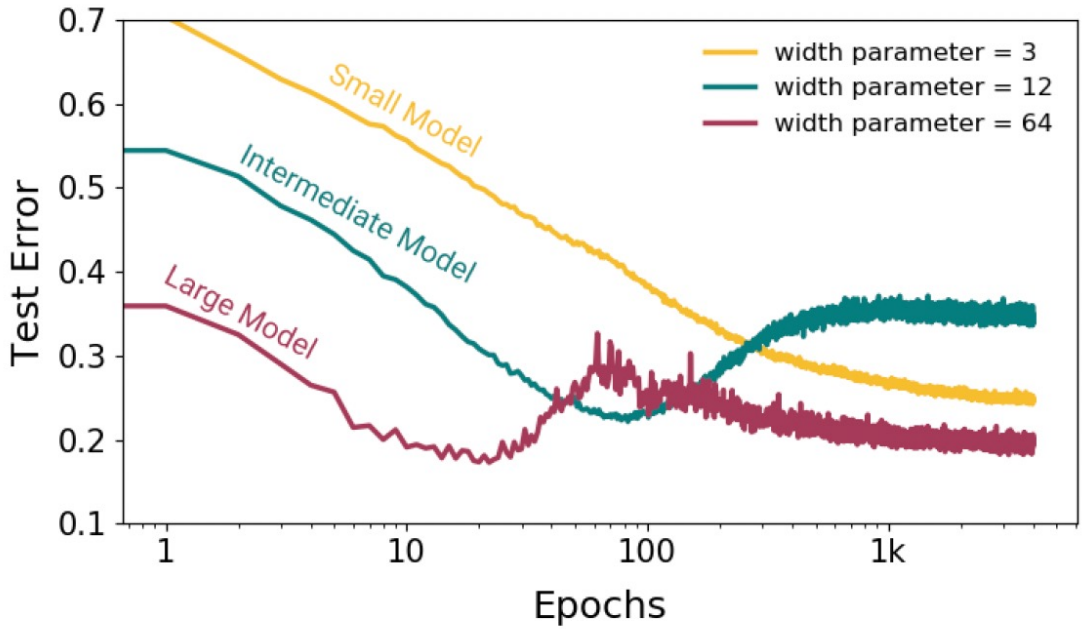
Ширина нейросети

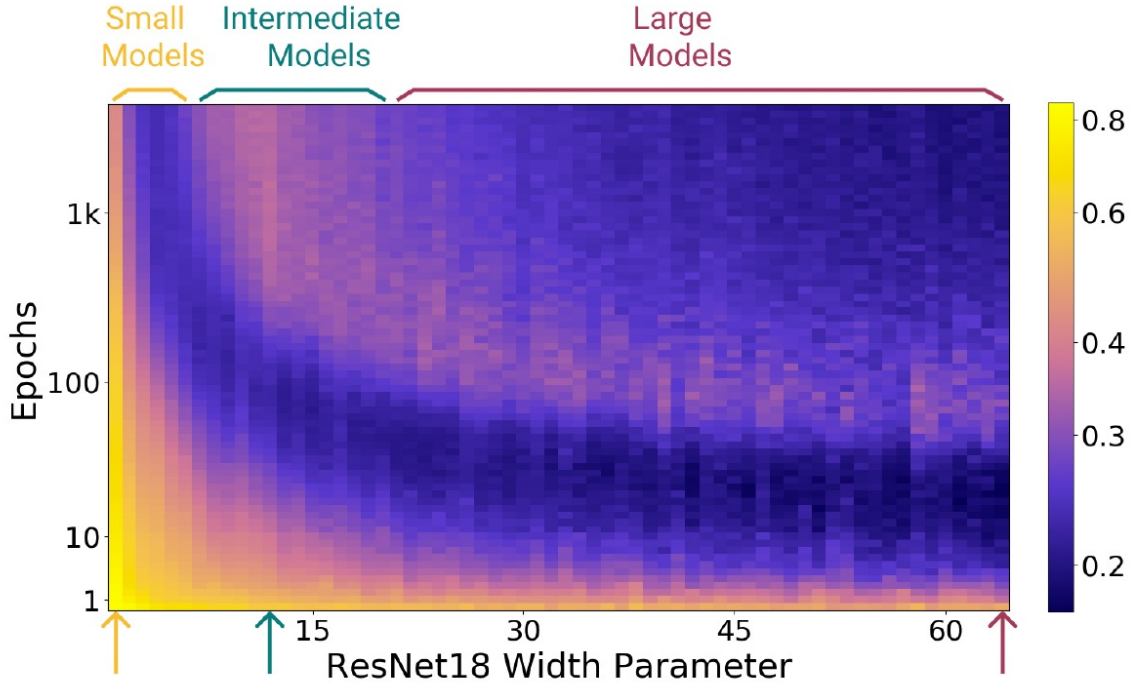


Ансамбли разрешающих деревьев

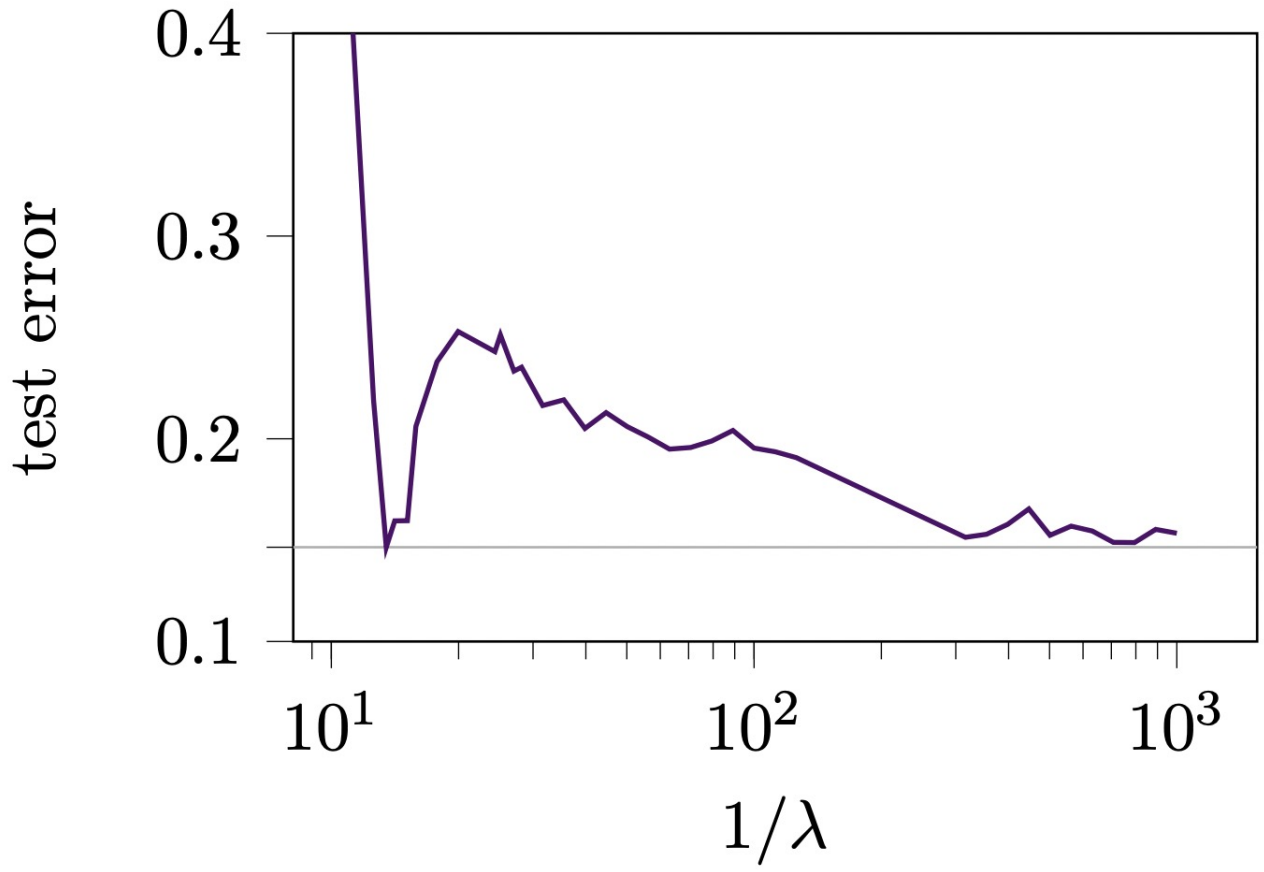


Число эпох обучения





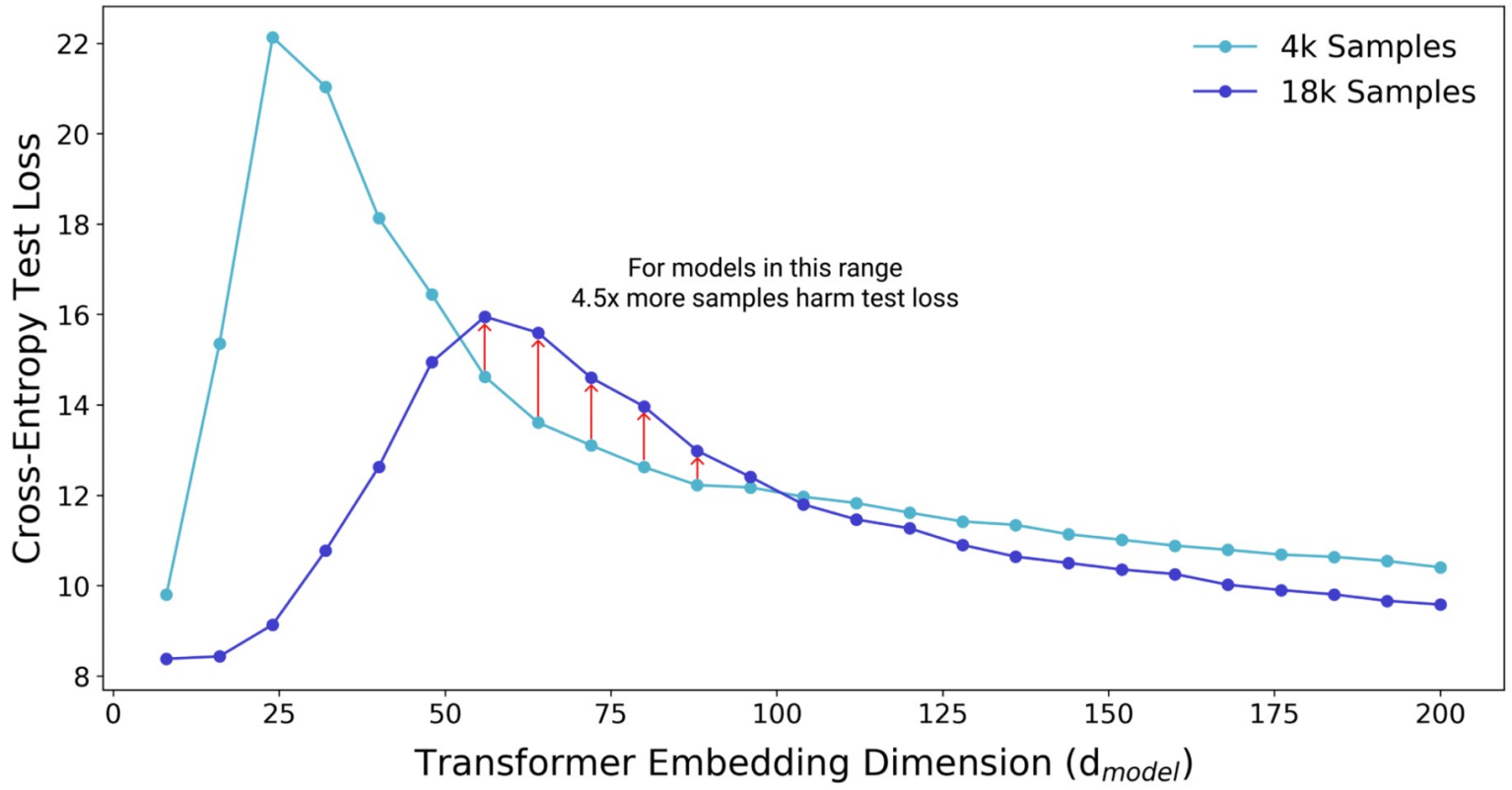
Коэффициент регуляризации



Проблемы

- Эффективную сложность модели очень сложно оценить, непонятно, в каком режиме мы находимся
- Можно заметить рост ошибки и прекратить увеличивать сложность модели, упустив возможность улучшить результат

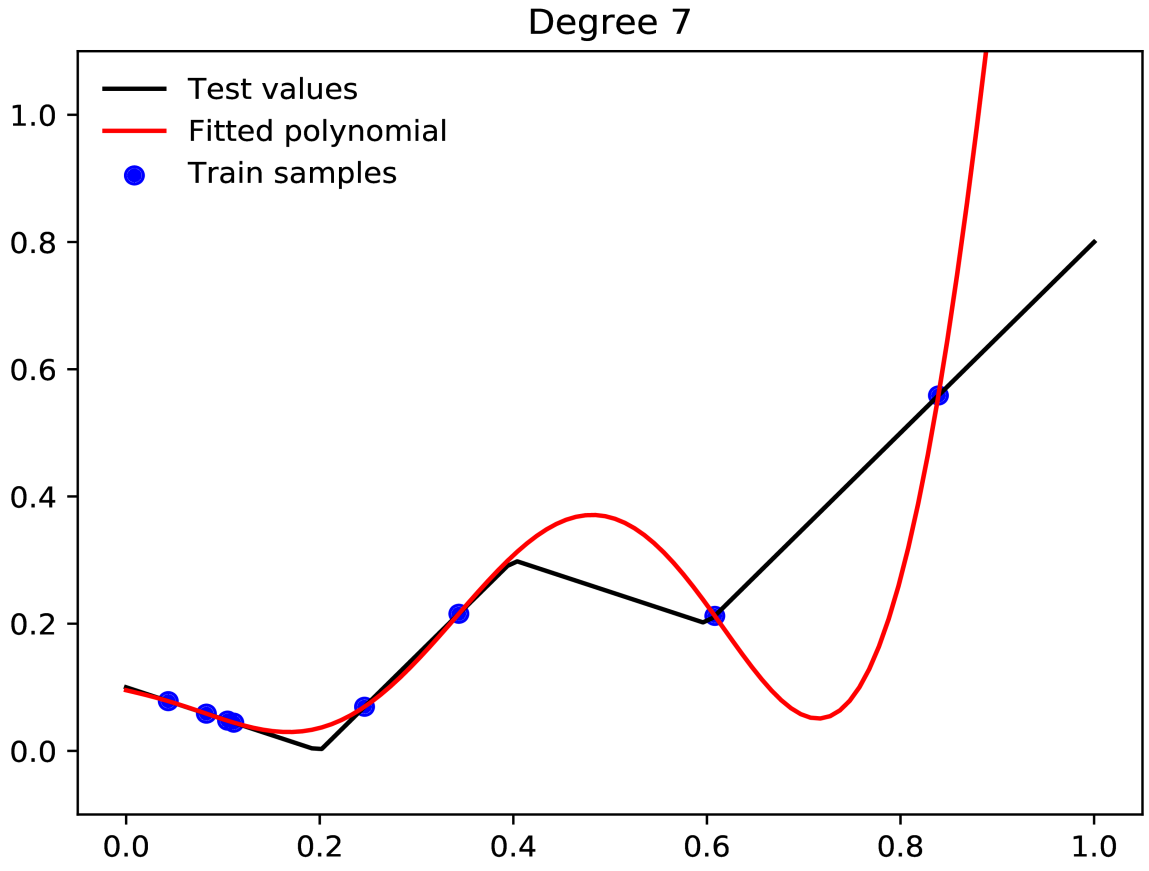
Проблемы: число данных



Почему это все происходит?

- Вообще говоря никто не знает
- Интуитивно понятно, что модели в районе EMC должны быть самыми плохими

Почему это все происходит?



Почему это все происходит?

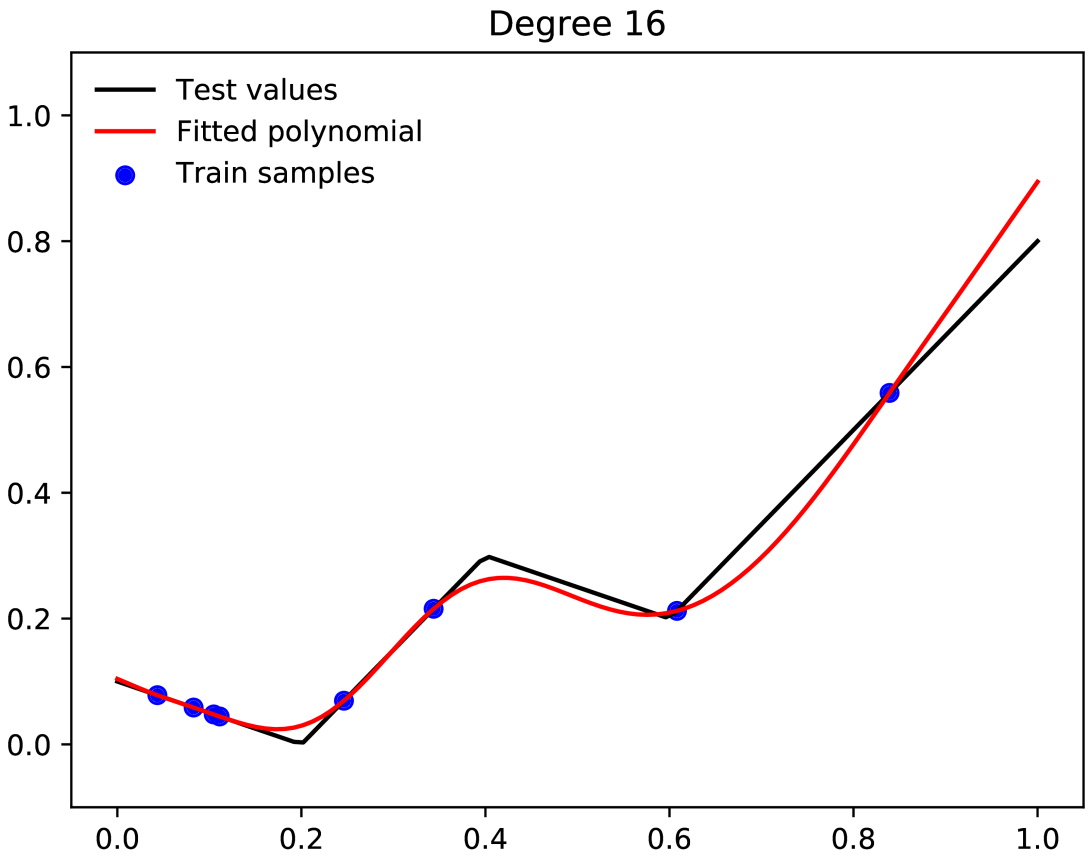
- Вообще говоря никто не знает
- Интуитивно понятно, что модели в районе EMC должны быть самыми плохими
- Есть теоретические обоснования для линейных моделей
- Есть эмпирические свидетельства для более сложных моделей

«Сложные» признаки

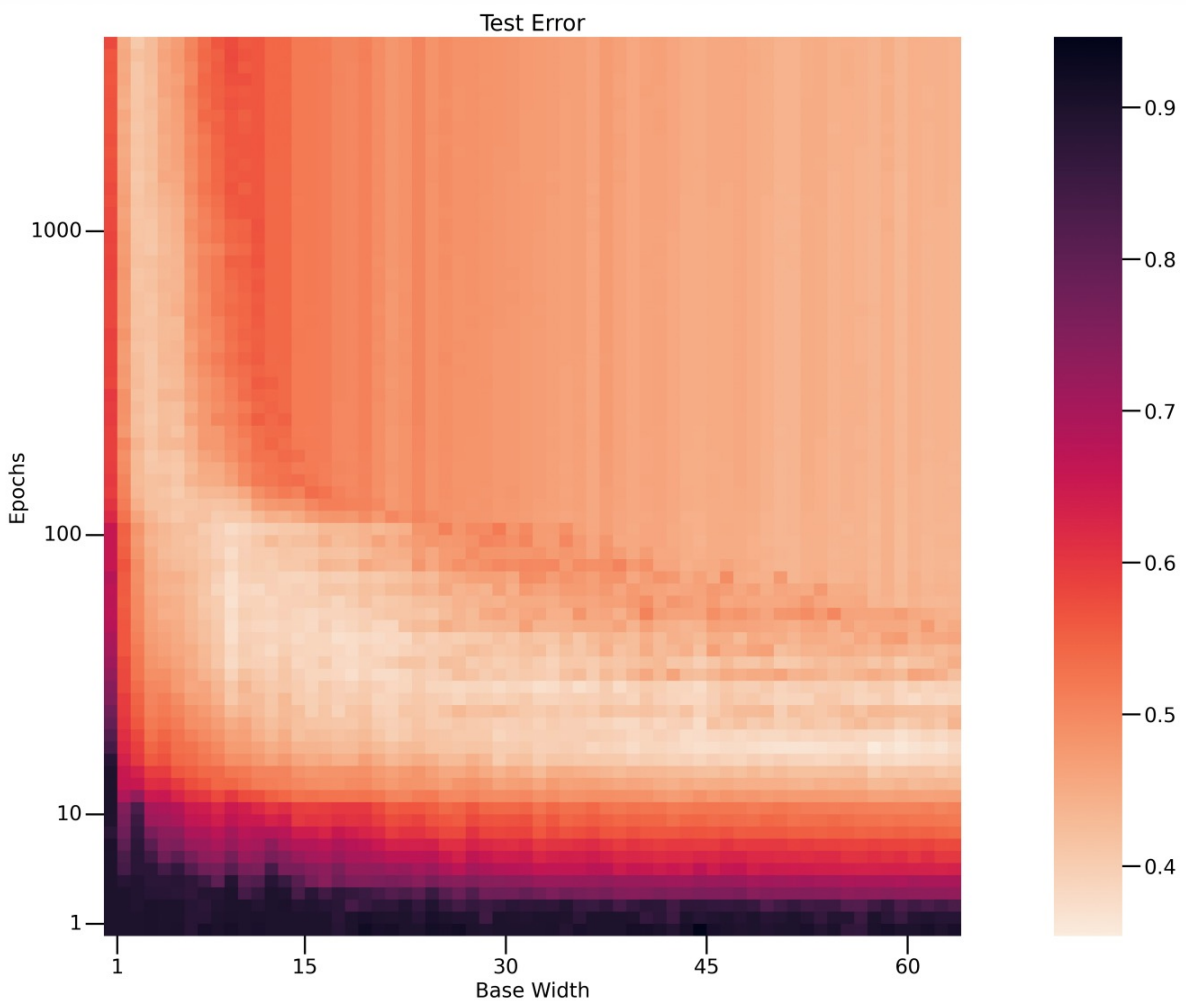
- Можно представить модель как сумму «быстро» и «медленно» обучаемых моделей
- Если «быстро» обучаемая модель переобучается быстрее, чем «медленная» начинает давать результат, будет двойной спуск

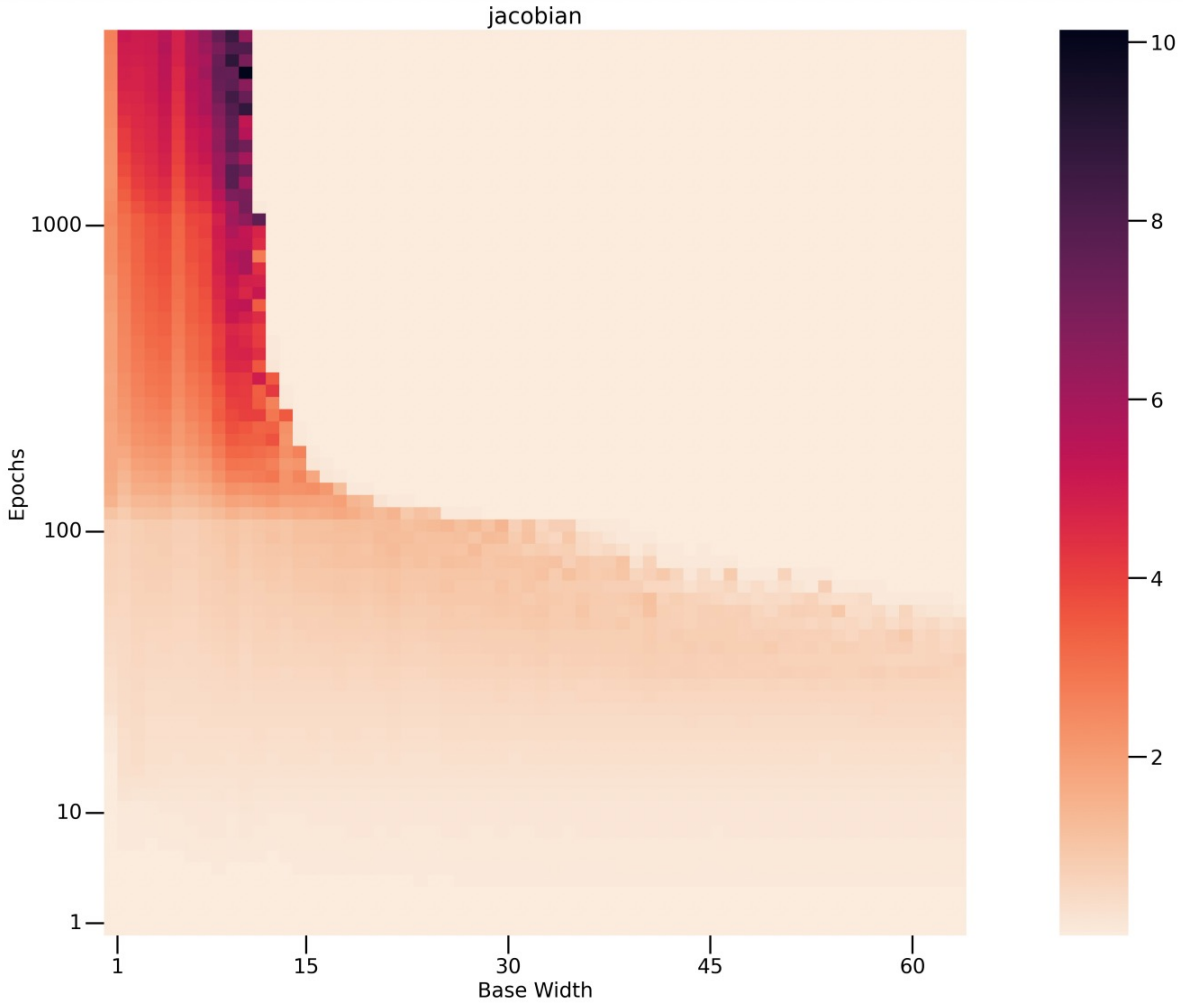
Гладкость

- Более сложные модели обычно более гладкие
- Настоящие результаты обычно тоже гладкие
- За счет этого в «современном» режиме из всех возможных результатов обучения обычно выбирается более близкий к реальности



Гладкость


Test Error
Epochs
1000
100
10
1
1
15
30
45
60
Base Width
0.9
0.8
0.7
0.6
0.5
0.4


jacobian
Epochs
1000
100
10
1
1
15
30
45
60
Base Width
10
8
6
4
2

Выводы

- Иногда модели разумно усложнять даже после того, как они вроде бы начали переобучаться
- Предсказать это довольно сложно
- Но такие модели могут давать более высокое качество, чем «обычные»

Источники

- «Deep Double Descent: Where Bigger Models and More Data Hurt» https://arxiv.org/abs/1912.02292
- «Reconciling modern machine learning practice and the bias-variance trade-off» https://arxiv.org/abs/1812.11118
- «Regularization-wise double descent: Why it occurs and how to eliminate it» https://arxiv.org/abs/2206.01378
- «Deep Double Descent via Smooth Interpolation» https://arxiv.org/abs/2209.10080
- «When and how epochwise double descent happens» https://arxiv.org/abs/2108.12006