Главная проблема машинного обучения – алгоритм должен хорошо работать на новых данных, которых он раньше не видел, а не только на тех, что использовались для обучения модели. Эта способность правильной работы на ранее не предъявлявшихся данных называется **обобщением**.

Ошибка обучения – мера ошибки на обучающем наборе данных

Ошибка обобщения – математическое ожидание ошибки.

Как правило, для оценки ошибки обобщения модели измеряется ее качество на **тестовом наборе** данных, отдельном от обучающего набора.

Факторы, определяющие качество работы алгоритма машинного обучения, :

1) сделать ошибку обучения как можно меньше;

2) сократить разрыв между ошибками обучения и **обобщения**.

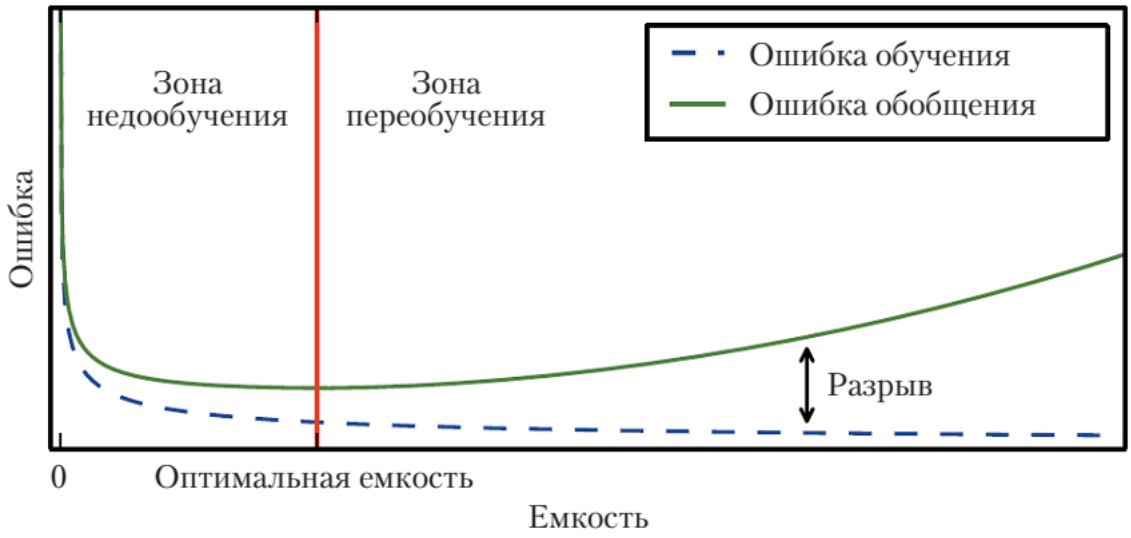


Рис. 5.3 ❖ Типичная связь между емкостью и ошибкой. Ошибки обучения и тестирования ведут себя по-разному. В левой части графика обе ошибки принимают большие значения. Это режим недообучения. По мере увеличения емкости ошибка обучения снижается, а разрыв между ошибкой обучения и обобщения растет. В конечном итоге величина разрыва перевешивает уменьшение ошибки обучения, и мы попадаем в режим переобучения, где емкость слишком сильно превышает оптимальную емкость

Эти факторы соответствуют двум центральным проблемам машинного обучения: **недообучению** и **переобучению**.

Емкость (сложность) (capacity) **модели описывает ее способность к аппроксимации широкого спектра функций**. Один из способов контроля над емкостью алгоритма обучения состоит в том, чтобы выбрать его **пространство гипотез** – множество функций, которые алгоритм может рассматривать в качестве потенциального решения.

Принцип экономии («бритва 'Оккама»): из всех гипотез, одинаково хорошо объясняющих наблюдения, следует выбирать «простейшую».

Эта идея была формализована и уточнена в XX веке основателями **теории статистического обучения** (Vapnik , Chervonenkis)

Общая постановка задачи обучения с учителем.

Модель обучения с учителем по эмпирическим данным предполагает наличие:

- генератора случайных **входных объектов** $\boldsymbol{x}$ – элементов некоторого пространства X, появляющихся независимо согласно фиксированному, но неизвестному распределению вероятностей $P(\boldsymbol{x})$;
- учителя, определяющего **выход** y – элемента некоторого пространства Y, для любого входящего $\boldsymbol{x}$, согласно условному распределению $P(y/\boldsymbol{x})$, также фиксированному и неизвестному;
- **класса функций** $\{f(\boldsymbol{x},\alpha)\}$, где параметр α может принимать значения из некоторого допустимого множества произвольной природы.

Задача обучения состоит в **выборе из заданного множества функций одной функции, которая предсказывает ответ учителя наилучшим образом**.

Этот выбор должен быть основан на **тренировочной последовательности** (ТП) конечного объема (l), т.е. независимых, одинаково распределенных согласно закону $P(\boldsymbol{x},y) = P(\boldsymbol{x})\times P(y/\boldsymbol{x})$ наблюдениях $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)$.

Наилучшая функция f, которую можно выбрать, – функция, минимизирующая **ожидаемый риск** (**ошибка обобщения**, generalization error)

$$R[f] = \int_{X\times Y} L(f(\boldsymbol{x}),y)dP(\boldsymbol{x},y),$$

где $L(\cdot)$ – содержательно обоснованная **функция потерь**.

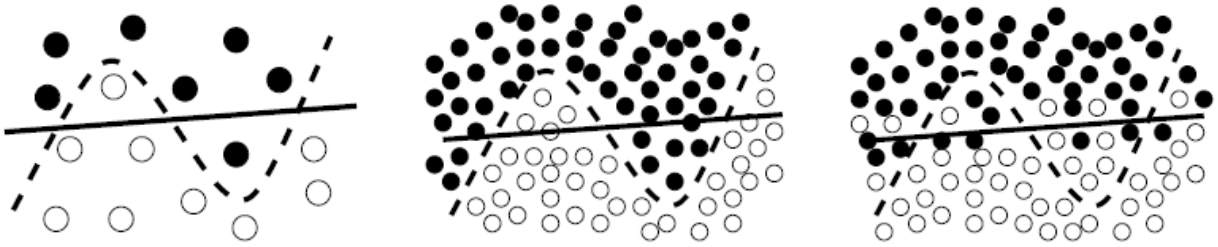
Поскольку распределение $P(\boldsymbol{x},y)$ неизвестно, можно, руководствуясь индукционным принципом минимизации **эмпирического риска**, заменить среднее по мере $P(\boldsymbol{x},y)$ средним по тренировочным данным:

$$R_{emp}[f] = \frac{1}{l}\sum_{i=1}^{l} L(f(\boldsymbol{x}_i),y_i)$$

эмпирический риск (**ошибка обучения**, training error)

Однако для конечных выборок этот принцип оказывается несостоятельным, поскольку может приводить к **переобучению** (overfitting).

Об эффекте переобучения говорят, если качество решающей функции вне ТП, на вновь поступающих образцах, оказывается существенно хуже качества, достигнутого на ТП.

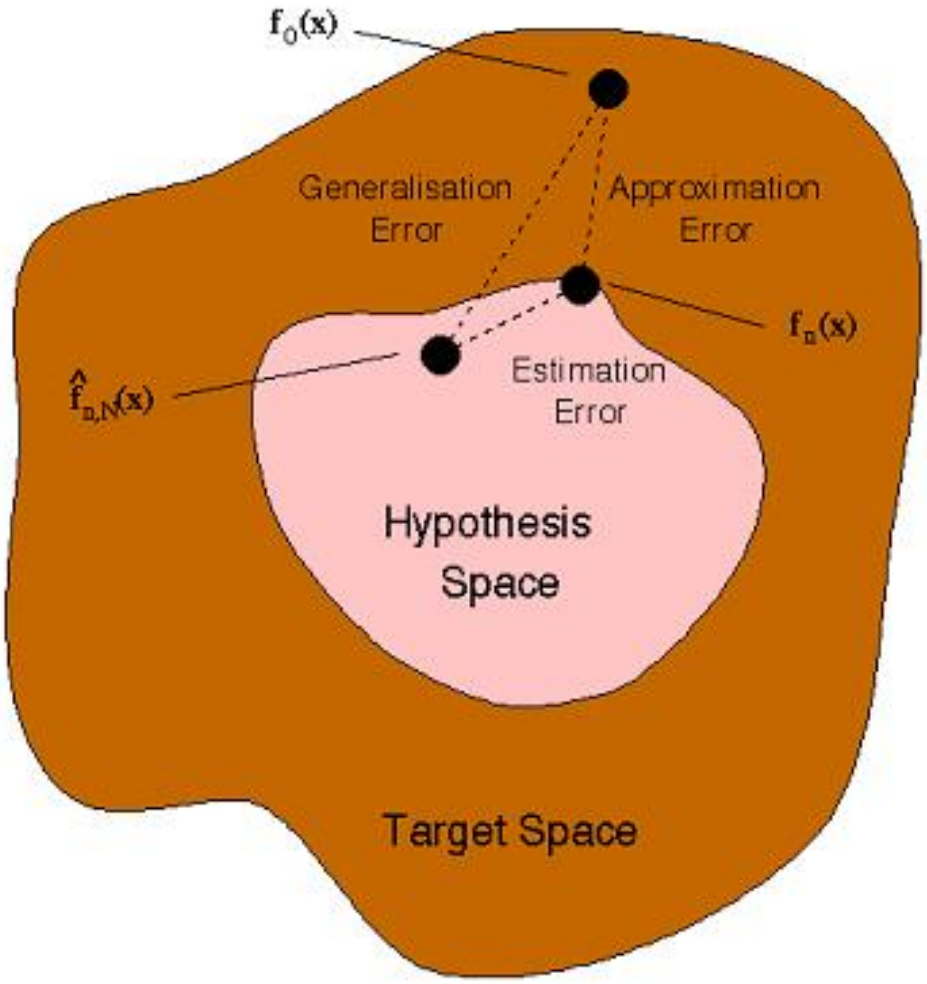


Структурная минимизация риска

Один из путей преодоления явления переобучения состоит в **сужении класса аппроксимирующих функций до класса со сложностью, подходящей для имеющейся тренировочной последовательности**.

(Мера сложности функций – мера способности их к обобщению)

[Один из способов контроля над емкостью алгоритма обучения состоит в том, чтобы выбрать его **пространство гипотез** – множество функций, которые алгоритм может рассматривать в качестве потенциального решения.]



Как найти **оптимальную сложность** класса функций для имеющегося объема ТП?

Статистическая теория обучения предлагает новый индукционный принцип для обучения по конечным выборкам – **структурную минимизацию риска** (structural risk minimization), который дает определенный способ, позволяющий контролировать сложность класса гипотез.

При условии, что эмпирический риск с ростом объема ТП по вероятности равномерно по α стремится к ожидаемому риску:

$$\lim_{l\to\infty} P(\sup_{\alpha} \left| R_{emp}[f] - R[f] \right| > \varepsilon) = 0 , \quad \forall \varepsilon > 0 ,$$

верхняя граница ожидаемого риска, основанная на некоторой **мере сложности функций *h***, имеет вид:

$$R[f] \le R_{emp}[f] + \Omega(h/l, \ln \eta / l) \qquad (1)$$

и гарантирована с вероятностью 1–η для любого $\eta \in (0,1)$ и любой функции *f* из заданного класса аппроксимирующих функций $\{f(\mathbf{x},\alpha)\}$ с мерой сложности *h* для $l > h$.

Второе слагаемое в (1) контролирует сложность класса аппроксимирующих функций и не зависит от ТП.

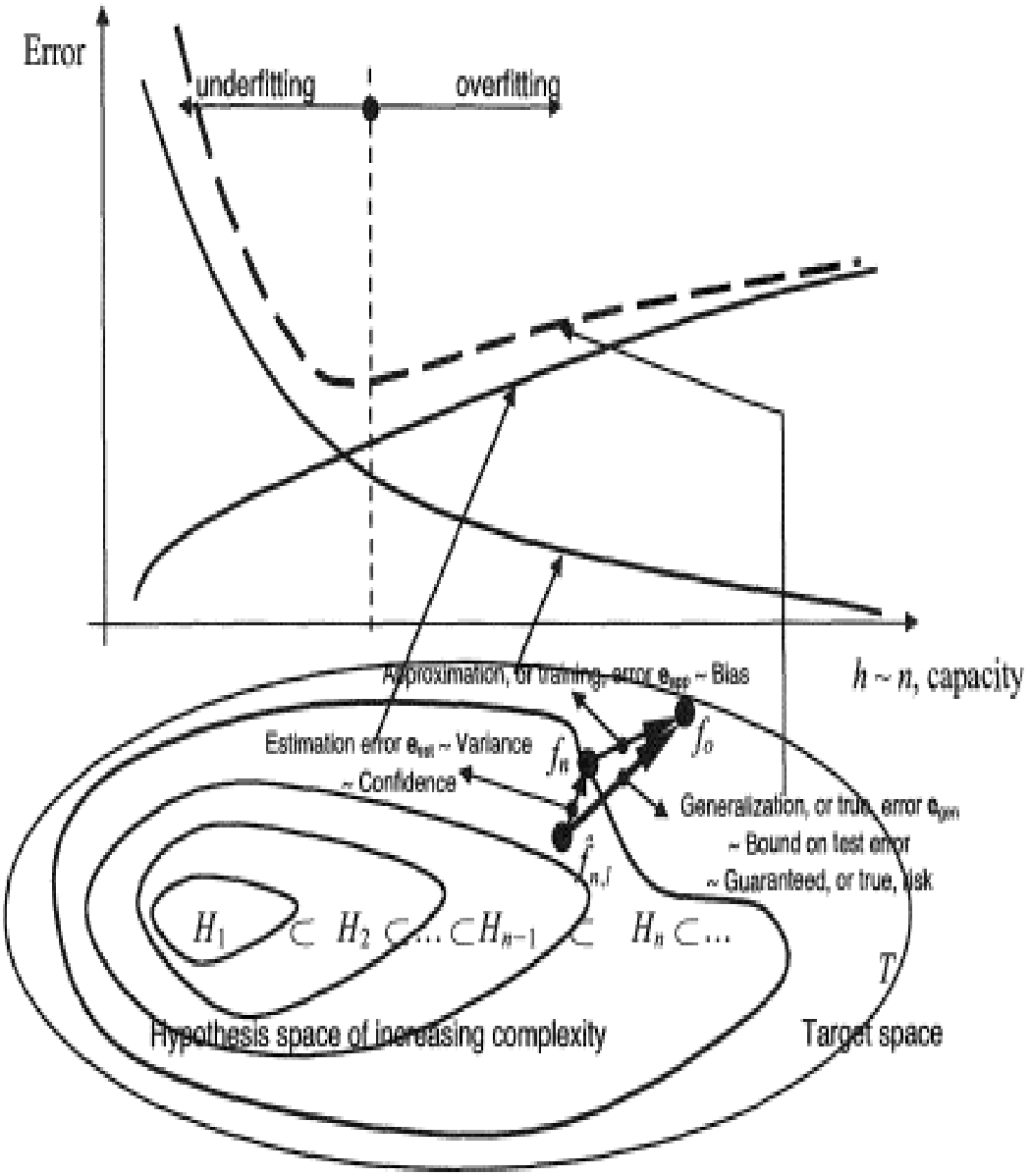
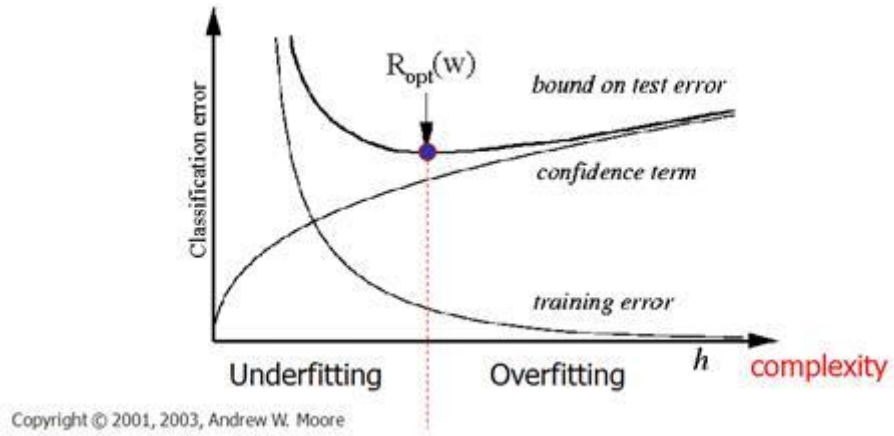


Figure 2.6
Structure of nested hypothesis functions and different errors that depend on the number of basis functions n for fixed sample size l. h denotes the VC dimension, which is equal to $n + 1$, for the linear combination of n fixed basis functions. Confidence is a confidence interval on the training error.

Принцип структурной минимизации риска предлагает определенный способ, позволяющий контролировать сложность класса функций, из которого выбирается решение. На заданном множестве гипотез $\{f(\boldsymbol{x},\alpha)\}$ вводится вложенная структура согласно некоторой мере сложности подклассов (h). На каждом элементе структуры выбирается функция, минимизирующая эмпирический риск. Затем из отобранных функций выбирается функция, доставляющая минимум верхней границе ожидаемого риска (1).



Сущность нового индукционного принципа состоит в минимизации верхней границы ожидаемого риска, что обеспечивает согласованность между **ошибкой обучения** и **сложностью класса аппроксимирующих функций**.

Принцип структурной минимизации риска оснащает SV-машины **способностью к обобщению**, которая и является целью статистического обучения.