3.6 Generalization

- **Motivation:** Not simply memorized seen data → Discovered a general pattern
- Irl, finite collection of data → the risk that might fit training data.
- Overfitting: fitting closer to traing data than to the underlying distribution

Solve ↗

Regularization methods

Model Complexity

→ If our model class was not capable of fitting arbitrary labels, then it must have discovered a pattern.

→ Weight decay: ur 1st regularization technique.

→ Error on the hold out data, validation error.

Underfitting or Overfitting

Must watch out:

(1) Generalization Gap ($R_{emp} - R$) small

→ Underfitting → Get more complex model

(2) $R_{emp} \ll$ validation error

→ Overfitting

Polynomial Curve fitting

$\hat{y} = \sum_{i=0}^{d} x^{(i)} w_i$ (linear regression problem)

w_i: weights

d: degree

w_0: bias since $x^0 = 1$

Training Error and Generalization Error

- IID assumption (Independently from Identical Distributions)

where: $P(X, Y) = Q(X, Y)$

(we believe) ↑ data sampled distribution ↑ test generated distribution

- Training Error R_{emp}: statistic calculated on training data set (finite, or samples)
- Generalization Error R: expectation taken with respect to underlying distribution (infinite stream)

matrix ↓ vector ↓

$$R_{emp}[\mathbf{X}, \mathbf{y}, f] = \frac{1}{n} \sum_{i=1}^{n} l(x^{(i)}, y^{(i)}, f(x^{(i)}))$$

$$R[p, f] = E_{(x,y) \sim P}[l(x, y, f(x))]$$

↑ func

$$= \iint l(x, y, f(x)) \cdot p(x, y)\, dx\, dy$$

p: PDF associated with P

P: a prob distribution

⇒ Estimate R, but R_{emp} be biased on training set

! When should we expect traing error be close to population error (R)?

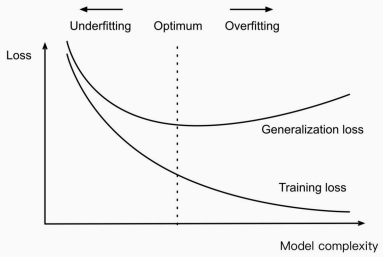

Fig. 3.6.1 Influence of model complexity on underfitting and overfitting.

Summary

- Rules of thumb

1. Use validation sets (or K-*fold cross-validation*) for model selection;
2. More complex models often require more data;
3. Relevant notions of complexity include both the number of parameters and the range of values that they are allowed to take;
4. Keeping all else equal, more data almost always leads to better generalization;
5. This entire talk of generalization is all predicated on the IID assumption. If we relax this assumption, allowing for distributions to shift between the train and testing periods, then we cannot say anything about generalization absent a further (perhaps milder) assumption.

Dataset Size

→ another big consideration

Model Selection

→ After evaluating multiple models

(different architectures, training objectives, selected features, data preprocessing, learning rates, etc.)

→ Never rely on test data for model selection

(Be honest!)

⇒ Data set

↙ ↓ ↘

Train Validation Test

Cross-Validation

↳ Employ K-fold cross-validation:

train K-1 subsets, validate on a different subset

→ The training, validation errors = average from K experiments