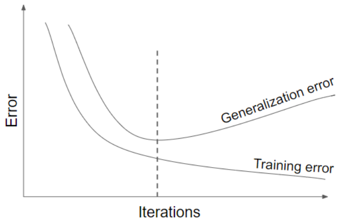
Introduction to Machine Learning

Early Stopping



Learning goals

- XXX
- XXX

EARLY STOPPING

- When training with an iterative optimizer such as SGD, it is commonly the case that, after a certain number of iterations, generalization error begins to increase even though training error continues to decrease.
- **Early stopping** refers to stopping the algorithm early before the generalization error increases.

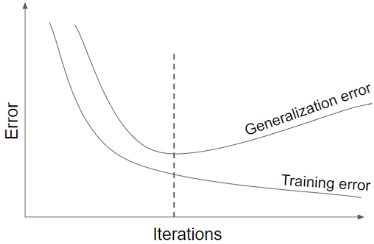


Figure: After a certain number of iterations, the algorithm begins to overfit.

EARLY STOPPING

How early stopping works:

1. Split training data $\mathcal{D}_{\text{train}}$ into $\mathcal{D}_{\text{subtrain}}$ and $\mathcal{D}_{\text{val}}$ (e.g. with a ratio of 2:1).
2. Train on $\mathcal{D}_{\text{subtrain}}$ and evaluate model using the validation set $\mathcal{D}_{\text{val}}$.
3. Stop training when validation error stops decreasing (after a range of “patience” steps).
4. Use parameters of the previous step for the actual model.

More sophisticated forms also apply cross-validation.

EARLY STOPPING

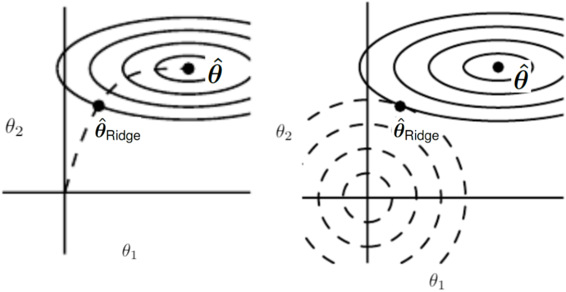
Strengths	Weaknesses
Effective and simple	Periodical evaluation of validation error
Applicable to almost any model without adjustment	Temporary copy of $\boldsymbol{\theta}$ (we have to save the whole model each time validation error improves)
Combinable with other regularization methods	Less data for training $\rightarrow$ include $\mathcal{D}_{\text{val}}$ afterwards

- Relation between optimal early-stopping iteration T_{stop} and weight-decay penalization parameter λ for step-size α (see Goodfellow et al. (2016) page 251-252 for proof):

$$T_{\text{stop}} \approx \frac{1}{\alpha\lambda} \Leftrightarrow \lambda \approx \frac{1}{T_{\text{stop}}\alpha}$$

- Small λ (low penalization) $\Rightarrow$ high T_{stop} (complex model / lots of updates).

EARLY STOPPING



Credit: Goodfellow et al. (2016)

Figure: An illustration of the effect of early stopping. *Left:* The solid contour lines indicate the contours of the negative log-likelihood. The dashed line indicates the trajectory taken by SGD beginning from the origin. Rather than stopping at the point $\hat{\boldsymbol{\theta}}$ that minimizes the risk, early stopping results in the trajectory stopping at an earlier point $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$. *Right:* An illustration of the effect of L_2 regularization for comparison. The dashed circles indicate the contours of the L_2 penalty which causes the minimum of the total cost to lie closer to the origin than the minimum of the unregularized cost.