LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Deep Learning

Chapter 3: Early Stopping

Bernd Bischl

Department of Statistics – LMU Munich

WS 2021/2022

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.

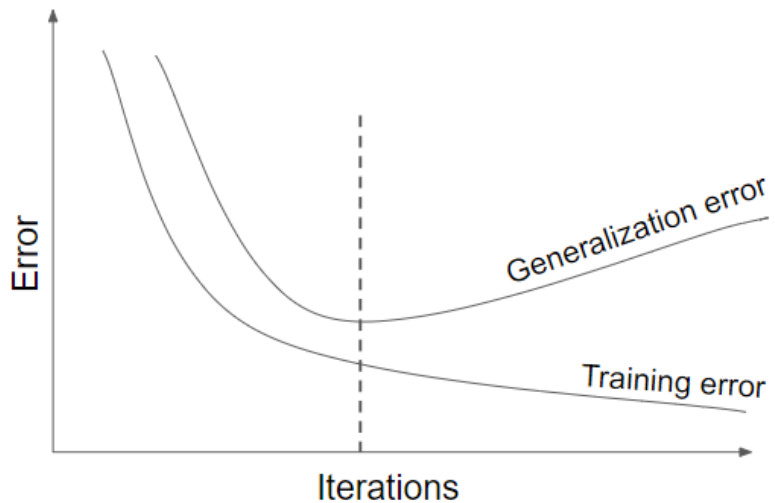


Figure: After a certain number of iterations, the algorithm begins to overfit.

EARLY STOPPING

Algorithm Early Stopping

Initalize $\boldsymbol{\theta}$ and set $\nu^* = \infty$
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)
while stop criterion not met: **do**
    Update $\boldsymbol{\theta}$ using $\mathcal{D}_{\text{subtrain}}$ for a predefined number of optimization steps
    Evaluate the model on $\mathcal{D}_{\text{val}}$ and save the resulting validation set error in ν
    if $\nu < \nu^*$: **then**
        $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$
        $\nu^* \leftarrow \nu$
    end if
end while
Return $\boldsymbol{\theta}^*$

- A possible stopping criterion is the maximum number of times to observe a worsening validation set error after $\boldsymbol{\theta}^*$ was updated.
- More sophisticated forms of early stopping 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 $\boldsymbol{\theta}$ as $\boldsymbol{\theta}^*$ each time the validation error improves)
Combinable with other regularization methods	Less data for training $\rightarrow$ include $\mathcal{D}_{\text{val}}$ afterwards

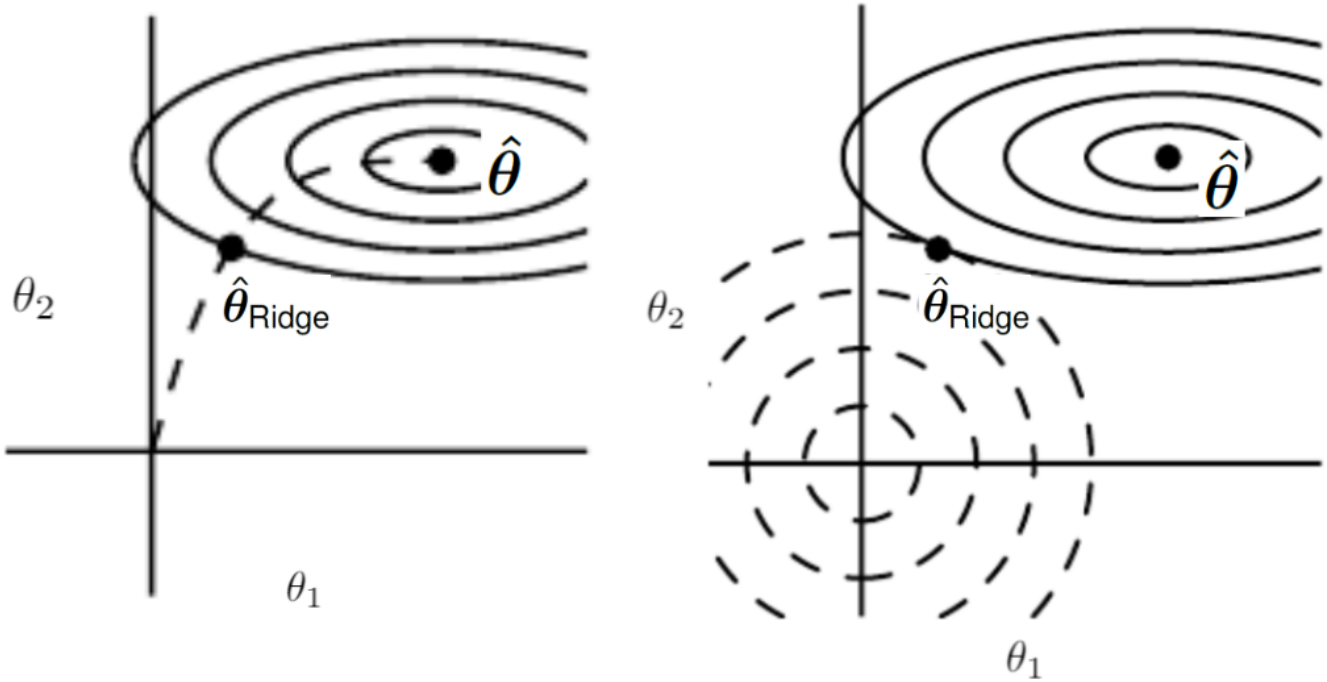
EARLY STOPPING

- Early stopping restricts the parameter space: Assuming the gradient is bounded, restricting both the number of iterations and the learning rate limits the volume of parameter space reachable from initial parameters.
- For a simple linear model and quadratic error function the following relation between optimal early stopping iteration T_{stop} and weight decay penalization parameter λ for learning rate α holds (see Goodfellow et al. (2016) ch. 7.8 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), ch. 7

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 L2 regularization for comparison. The dashed circles indicate the contours of the L2 penalty, which causes the minimum of the total cost to lie nearer the origin than the minimum of the unregularized cost.