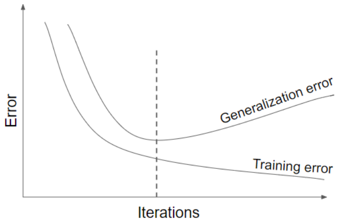
Introduction to Machine Learning

Regularization
Early Stopping

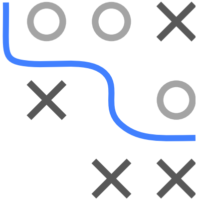


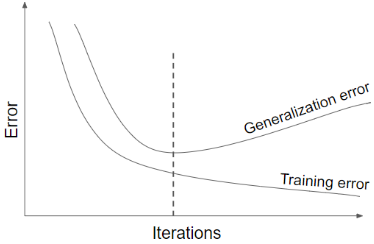
Learning goals

- Know how early stopping works
- Understand how early stopping acts as a regularizer

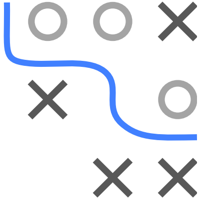
EARLY STOPPING

- Especially for complex nonlinear models we can easily overfit
- In optimization: Often, after a certain number of iterations, generalization error begins to increase even though training error continues to decrease





EARLY STOPPING / 2

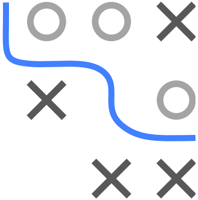


For iterative optimizers like SGD,
we can monitor this step-by-step over small iterations:

1. Split train data $\mathcal{D}_{\text{train}}$ into $\mathcal{D}_{\text{subtrain}}$ and $\mathcal{D}_{\text{val}}$ (e.g. with ratio of 2:1)
2. Train on $\mathcal{D}_{\text{subtrain}}$ and eval model on $\mathcal{D}_{\text{val}}$
3. Stop 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 AND $L2$ ▸ Goodfellow, Bengio, and Courville 2016



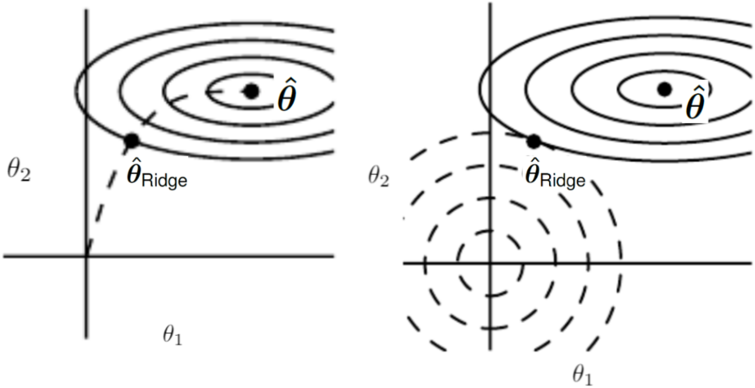
Strengths	Weaknesses
Effective and simple	Periodical evaluation of validation error
Applicable to almost any model without adjustment	Temporary copy of $\boldsymbol{\theta}$ (we have to save the whole model each time validation error improves)
Combinable with other regularization methods	Less data for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

- For simple case of LM with squared loss and GD optim initialized at $\boldsymbol{\theta} = 0$: Early stopping has exact correspondence with $L2$ regularization/WD: optimal early-stopping iter T_{stop} inversely proportional to λ scaled by step-size α

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

- Small λ (regu. $\downarrow$) $\Rightarrow$ large T_{stop} (complexity $\uparrow$) and vice versa

EARLY STOPPING AND $L2$ ▸ Goodfellow, Bengio, and Courville 2016 / 2



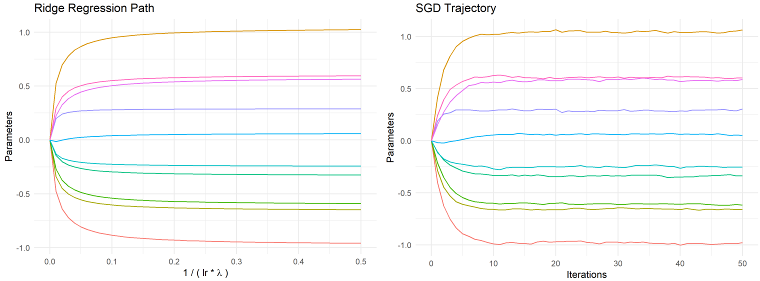
Goodfellow et al. (2016)

- Solid lines are $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$
- LHS: Trajectory of GD early stopped, initialized at origin
- RHS: Constrained form of ridge regularization

SGD TRAJECTORY AND *L2*

▸ Ali, Dobriban, and Tibshirani 2020

Solution paths for *L2* regularized linear model closely matches SGD trajectory of unregularized LM initialized at $\boldsymbol{\theta} = 0$



Caveat: Initialization at the origin is crucial for this equivalence to hold, which is almost never exactly used in practice in ML/DL applications