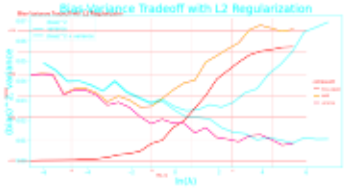
Introduction to Machine Learning

Regularization
Intuition for L2 Regularization in Non-Linear Models

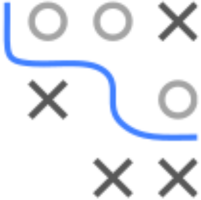


Learning goals

- Understand how regularization and parameter shrinkage can be beneficial to non-linear models

COUNTEREXAMPLE

Chris: I think ChatGPT produced a lot of "almost correct" stuff that culminated in a globally useless derivation. A general proof for DNNs imo can not work by giving a simple counterexample.



- A diagonal linear network with one hidden layer and one output unit can be written as $f(x|\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u} \odot \boldsymbol{v})^\top \boldsymbol{x}$
- optimizing the network with $L2$ regularization λ and MSE loss has multiple global minima that coincide with the lasso solution for the collapsed parameter $\boldsymbol{\theta} := \boldsymbol{u} \odot \boldsymbol{v}$ using 2λ
- Since there is no existence theorem (of a λ^* that reduces the MSE over OLS) for lasso compared to ridge regression, there can not be one for L2 regularized DNNs in general.

- Neyshabur et al., 2015 derive equivalent optimization problems for $L2$ regularized shallow relu-networks:

$$\underset{\boldsymbol{v}\in\mathbb{R}^H,(\boldsymbol{u}_h)_{h=1}^H}{\operatorname{argmin}} \left(\sum_{t=1}^{n} L\left(y_t, \sum_{h=1}^{H} v_h \left[\langle \boldsymbol{u}_h, \boldsymbol{x}_t\rangle\right]_+ \right) + \frac{\lambda}{2}\sum_{h=1}^{H}\left(\|\boldsymbol{u}_h\|^2 + |v_h|^2\right)\right),$$

is the same as

$$\underset{\boldsymbol{v}\in\mathbb{R}^H,(\boldsymbol{u}_h)_{h=1}^H}{\operatorname{argmin}} \left(\sum_{t=1}^{n} L\left(y_t, \sum_{h=1}^{H} v_h \left[\langle \boldsymbol{u}_h, \boldsymbol{x}_t\rangle\right]_+ \right) + \lambda\sum_{h=1}^{H}|v_h|\right),$$
$$\text{subject to } \|\boldsymbol{u}_h\| \leq 1 \quad (h = 1, \ldots, H).$$

- How can we do a general analysis of the effect of $L2$ regularization in DNNs when there are these close connections to other regularized problems for which there is no anaysis of the bias-variance trade-off and no existence theorem of an optimal $\lambda^* > 0$?