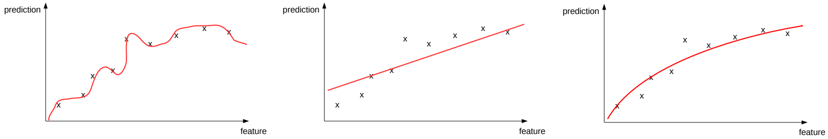
Deep Learning basics

Regularization

Learning goals

- Understand the concept of regularization
- Understand L2 regularization in more detail

REGULARIZATION

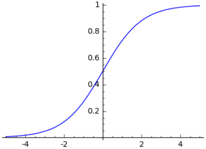


- Overfitting vs. underfitting
- Regularization: Any modification to a learning algorithm for reducing its generalization error but not its training error
- Build a preference into ML algorithm for one solution in hypothesis space over another
- Solution space is still the same
- Unpreferred solution is penalized: only chosen if there it fits training data much better

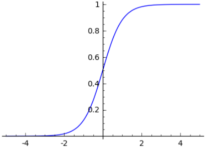
L2-REGULARIZATION

- Large parameters $\rightarrow$ overfitting

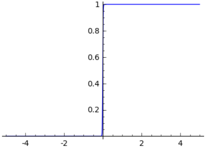
$\sigma(x)$



$\sigma(2x)$



$\sigma(100x)$



- Prefer models with smaller feature weights.
- Popular regularizers:
  - Penalize large L2 norm.
  - Penalize large L1 norm (aka LASSO, induces sparsity)

REGULARIZATION

- Add term that penalizes large l2 norm.
- The amount of penalty is controlled by a parameter λ
  - Linear regression:

$$J(\vec{\theta}) = MSE(\vec{\theta}) + \frac{\lambda}{2}\vec{\theta}^T\vec{\theta}$$

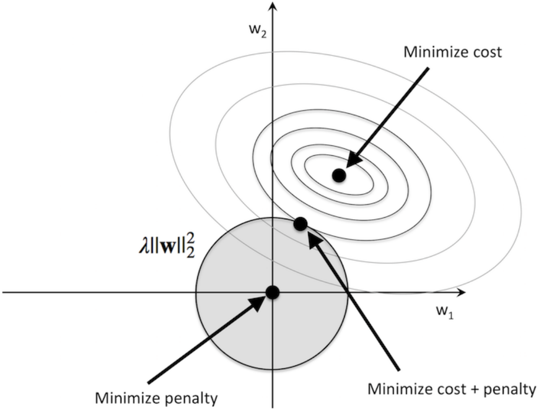
  - Logistic regression:

$$J(\vec{\theta}) = NLL(\vec{\theta}) + \frac{\lambda}{2}\vec{\theta}^T\vec{\theta}$$

- From a Bayesian perspective, l2-regularization corresponds to a Gaussian prior on the parameters.

L2-REGULARIZATION

- The surface of the objective function is now a combination of the original cost, and the regularization penalty.



L2-REGULARIZATION

- Gradient of regularization term:

$$\nabla_{\vec{\theta}} \frac{\lambda}{2} \vec{\theta}^T \vec{\theta} = \lambda \vec{\theta}$$

- Gradient descent for regularized cost function:

$$\vec{\theta}_{t+1} := \vec{\theta}_t - \eta \nabla_{\vec{\theta}} (NLL(\vec{\theta}_t) + \lambda \vec{\theta}_t^T \vec{\theta}_t)$$

$$\Leftrightarrow$$

$$\vec{\theta}_{t+1} := (1 - \eta\lambda) \vec{\theta}_t - \eta \nabla_{\vec{\theta}} NLL(\vec{\theta}_t)$$