Lecture 12: Regularization

Presentation · July 2019

1 author:

Alaa Tharwat
Fachhochschule Bielefeld

Lecture 12: Regularization

Alaa Tharwat

Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

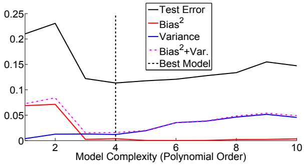
Lecture 12: Regularization

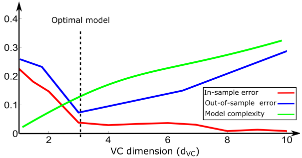
- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

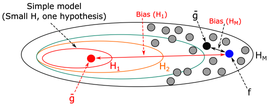
- The problem of overfitting occurs when we try to fit the data to the extent that the model fits the noise and outliers
- There are two types of noise, stochastic and deterministic

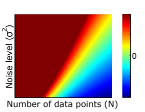
Impact of noise

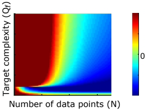
- Number of samples ↑ Overfitting ↓
- Stochastic noise ↑ Overfitting ↑
- Deterministic noise ↑ Overfitting ↑











Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

- Regularization technique is used for solving the overfitting problem by adding an extra term to the cost function
- The extra term is called the $\boxed{\textit{regularization term}}$ and it is used for handicapping the minimization of E_{in}, i.e. adding brakes on the model

$$C = C_0 + \text{regularization term}$$

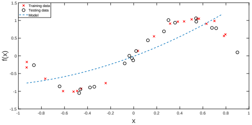
$$C = C_0 + \frac{\lambda}{2n} \sum_{\mathbf{w}} \mathbf{w}^2$$

where

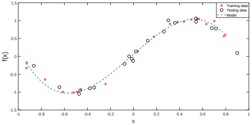
- The term $\frac{\lambda}{2n} \sum_{\mathbf{w}} \mathbf{w}^2$ is the sum of squares of all weights scaled by the factor $\frac{\lambda}{2n}$, and $\lambda > 0$ is called the *regularization parameter*.

- The regularization term aims to make a balance between minimizing the original cost function and finding small weights
- With a small λ, the original cost is minimized, but with a large λ, the weights are minimized
- **Side effect**: if we cannot fit the noise, maybe we cannot fit the target function (f)?

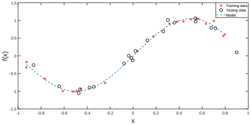
Example: Learning model with different complexities



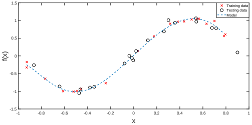
(a) Power=2, $E_{in} = 3.365$, and $E_{out} = 4.833$



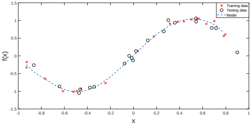
(b) Power=3, $E_{in} = 0.108$, and $E_{out} = 0.811$



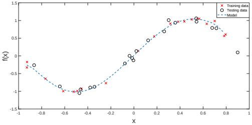
(c) Power=5, $E_{in} = 0.062$, and $E_{out} = 0.811$



(d) Power=9, $E_{in} = 0.062$, and $E_{out} = 0.914$



(e) Power=15, $E_{in} = 0.057$, and $E_{out} = 5.421$

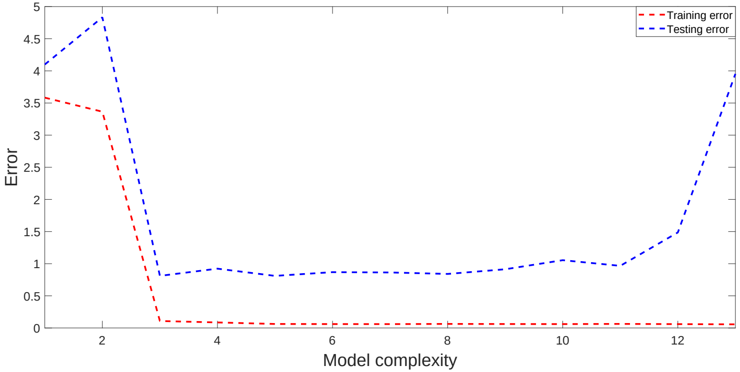


(f) Power=21, $E_{in} = 0.056$, and $E_{out} = 14.329$

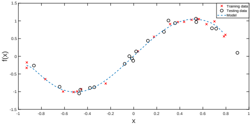
Figure: Visualization of the polynomial regression with different degrees.

From this example:

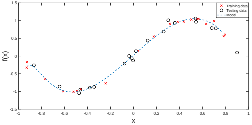
- Increasing the complexity of the model reduces E_{in} and increases E_{out}



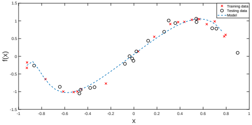
Example: Learning model with different regularization parameters



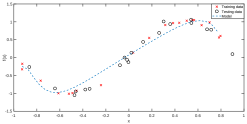
(a) $\lambda = 0.0001$,$E_{in} = 0.058$, and $E_{out} = 10.946$



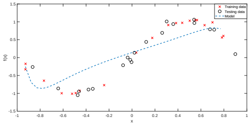
(b) $\lambda = 0.01$,$E_{in} = 0.069$, and $E_{out} = 2.185$



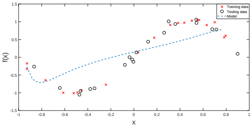
(c) $\lambda = 0.1$,$E_{in} = 0.133$, and $E_{out} = 1.158$



(d) $\lambda = 0.5$,$E_{in} = 0.367$, and $E_{out} = 4.060$



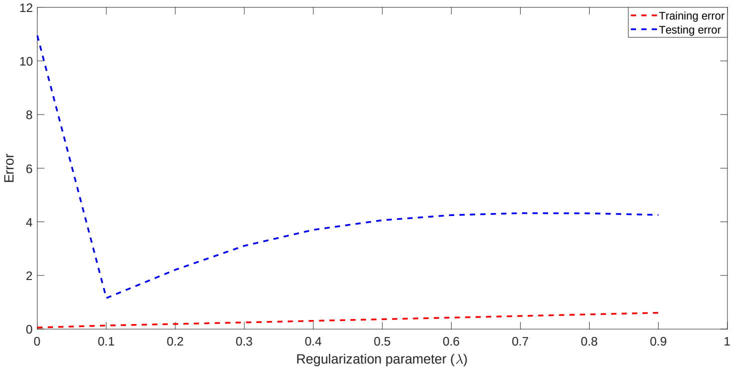
(e) $\lambda = 5$,$E_{in} = 2.457$, and $E_{out} = 2.118$



(f) $\lambda = 10$,$E_{in} = 2.457$, and $E_{out} = 2.118$

Figure: Visualization of the polynomial regression with different values of regularization parameter (power=21).

- Increasing λ increases the bias (side effect) slightly and reduces the testing error dramatically
- With regularization, the testing error reduced and hence the variance of the model becomes lower
- Large λ may lead to a simple model with high bias and high testing error



Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

- Given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ data samples transformed to $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_N, y_N)$

The unconstrained problem is

$$\begin{aligned} \text{minimize } E_{in}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{z}_n - y_n)^2 \\ &= \frac{1}{N}(\mathbf{Z}\mathbf{w} - Y)^T(\mathbf{Z}\mathbf{w} - Y) \end{aligned}$$

The unconstrained solution is $(\mathbf{w}_{lin})$ [Lecture 3]

$$\mathbf{w}_{lin} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T Y$$

- We can assume that a simple model (H_2) is a constrained version of a complex model (H_{10}).

$$h(\mathbf{x}) \in H_{10} = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_{10}\phi_{10}(x)$$

$$h(\mathbf{x}) \in H_2 = \underbrace{w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_{10}\phi_{10}(x)}_{\text{where } w_3 = w_4 = \cdots + w_{10} = 0}$$

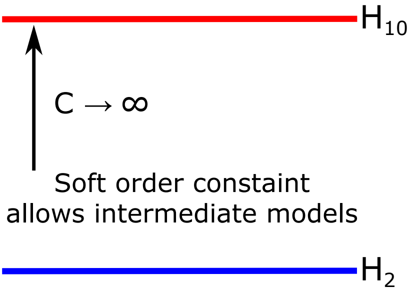
This is can be interpreted as $w_q = 0$ for $q > 2$ and H_2 is a constrained version of H_{10}, ($H_2 \subset H_{10}$)

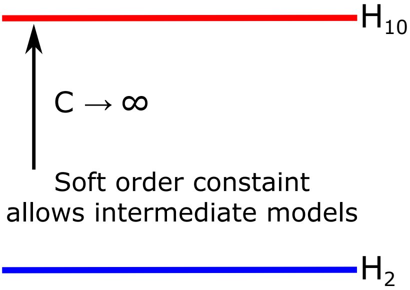
$$h(\mathbf{x}) \in H_C = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_{10}\phi_{10}(x)$$

such that: $\sum_{q=0}^{10} w_q^2 \leq C$

Softer version $\sum_{q=0}^{Q} w_q^2 \leq C$ softer-order constraint

- H_C is smaller/simpler than H_{10} $\Rightarrow$ better generalization



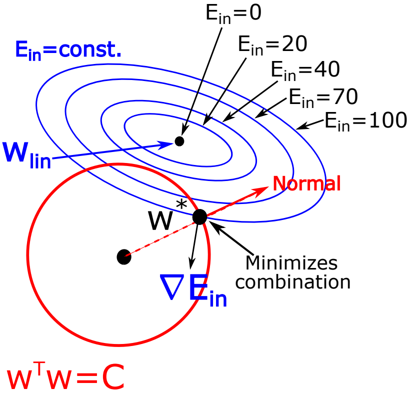


The parameter C puts a constraint on some weights to be small or zero (not exclude any order but gives it different weights)

$$\text{minimize: } \frac{1}{N}(\mathbf{Zw} - Y)^T(\mathbf{Zw} - Y)$$

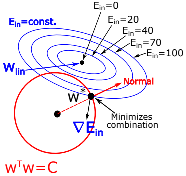
$$\text{subject to: } \mathbf{w}^T\mathbf{w} \leq C, \quad \Rightarrow \mathbf{w}_{reg} \in H_C \text{ instead of } \mathbf{w}_{lin}$$

- Surface $\mathbf{w}^T\mathbf{w} = C$, at optimal $\mathbf{w}$, should be perpendicular to $\bigtriangledown E_{in}$, otherwise; can move along the surface and decrease E_{in}
- Moving around the red circle changes the value of E_{in}
- Increasing the radius of the red circle (increase C) may be big and then include $\mathbf{w}_{lin}$ inside the circle and $\mathbf{w}_{lin}$ is the solution



$$\text{minimize: } \frac{1}{N}(\mathbf{Zw} - Y)^T(\mathbf{Zw} - Y)$$

$$\text{subject to: } \mathbf{w}^T\mathbf{w} \leq C$$



- Surface $\mathbf{ww}^T$ is $\perp \bigtriangledown E_{in}$, surface is $\perp$ normal. Hence, $\bigtriangledown E_{in}$ is parallel to normal, but, in opposite direction

$$\bigtriangledown E_{in}(\mathbf{w}_{reg}) \propto -\,\mathbf{w}_{reg}$$

$$= -2\frac{\lambda}{N}\mathbf{w}_{reg} \quad (\lambda \text{ is the Lagrange multiplier})$$

$$\bigtriangledown E_{in}(\mathbf{w}_{reg}) + 2\frac{\lambda}{N}\mathbf{w}_{reg} = 0$$

This is the differentiation of

$$\text{Min } E_{in}(w) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w} \qquad \boxed{C \uparrow \lambda \downarrow}$$

- With large $C \Rightarrow \lambda \approx 0$, $\mathbf{w}_{lin}$ is the solution, just minimize E_{in} as if there is no constraint
- With small $C \Rightarrow \lambda \uparrow$ and the regularization is more severe
- If $C = 0 \Rightarrow \lambda = \infty$ and $\mathbf{w} \approx 0$

The augmented error

$$
\begin{aligned}
\text{Min } E_{aug}(w) &= E_{in} + \text{regularization term} \\
&= E_{in}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w} \\
&= \frac{1}{N}(\mathbf{Zw} - Y)^T(\mathbf{Zw} - Y) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}
\end{aligned}
$$

It is similar to

$$
\text{Min } E_{in}(\mathbf{w}) = \frac{1}{N}(\mathbf{Zw} - Y)^T(\mathbf{Zw} - Y)
$$

$$
\text{Subject to : } \mathbf{w}^T\mathbf{w} \leq C \rightarrow \text{VC formulation}
$$

- This term $\mathbf{w}^T\mathbf{w} \leq C$ lends itself to the VC analysis because the hypotheses set is restricted (i.e. there are certain hypotheses that are no longer allowed) and a subset of the hypotheses is used and hence we expect a good generalization

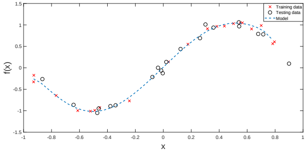
$$\text{Min } E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$
$$= \frac{1}{N}((\mathbf{Z}\mathbf{w} - Y)^T(\mathbf{Z}\mathbf{w} - Y) + \lambda\mathbf{w}^T\mathbf{w})$$

$$\bigtriangledown E_{aug}(\mathbf{w}) = 0 \Rightarrow \mathbf{Z}^T(\mathbf{Z}\mathbf{w} - y) + \lambda\mathbf{w} = 0$$

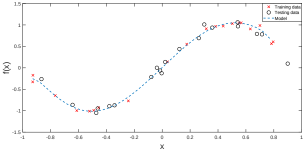
$$\mathbf{w}_{reg} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T\lambda \text{ (With regularization)}$$

- With a very large λ the term $\lambda\mathbf{I}$ dominates the $\mathbf{Z}^T\mathbf{Z}$ and the result of this term $(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}$ will be $\approx \frac{1}{\lambda}$ and hence $\mathbf{w}_{reg} \approx 0$

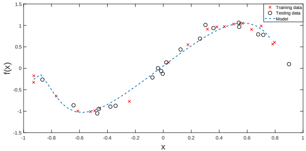
$$\lambda = 0 \Rightarrow \mathbf{w}_{lin} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\lambda \text{ (Without regularization)}$$



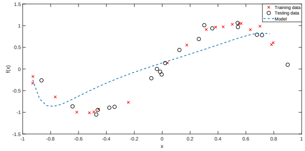
(a) $\lambda = 0$,$E_{in} = 0.056$, and $E_{out} = 14.329$



(b) $\lambda = 0.0001$,$E_{in} = 0.058$, and $E_{out} = 10.946$



(c) $\lambda = 0.1$,$E_{in} = 0.133$, and $E_{out} = 1.158$



(d) $\lambda = 5$,$E_{in} = 2.457$, and $E_{out} = 2.118$

- With $\lambda = 0$, this means that there is no regularization and this may lead to the overfitting problem
- Increasing λ relaxes the model and this may lead to the underfitting

In neural networks, the weights are updated. The partial derivatives after adding the regularization term are

$$\frac{\partial C}{\partial \mathbf{w}} = \frac{\partial C_0}{\partial \mathbf{w}} + \frac{\lambda}{n}\mathbf{w}$$
$$\frac{\partial C}{\partial b} = \frac{\partial C_0}{\partial b}$$

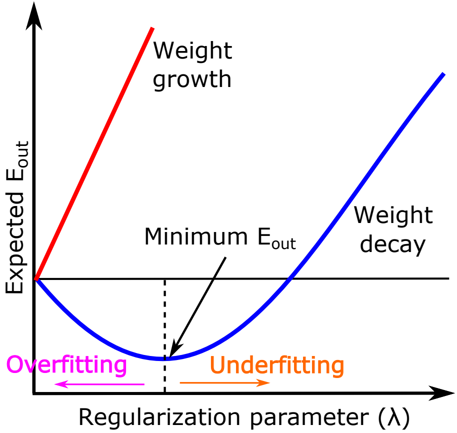
$$\begin{aligned}\mathbf{w}' &= \mathbf{w} - \eta\frac{\partial C_0}{\partial \mathbf{w}} - \frac{\eta\lambda\mathbf{w}}{n} \\ &= (1 - \frac{\eta\lambda}{n})\mathbf{w} - \eta\frac{\partial C_0}{\partial \mathbf{w}}\end{aligned}$$

where

- $\lambda = 0 \Rightarrow$ there is no regularization (original case)
- $\lambda > 0 \Rightarrow$ from the term $(1 - \frac{\eta\lambda}{n})$, the weights will be reduced, this is called $\boxed{\textit{weight decay}}$

Weight decay vs. weight growth

- In the weight growth, we constrain the weights to be large



- Stochastic noise is high frequency, and deterministic noise is also non-smooth
- The goal for any model is to constrain the model towards smoother hypotheses, why?
- Because the regularizer punishing the noise more than punishing the original signal/data

Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

- The regularizer is defined as follows, $\Omega = \Omega(h)$, the regularizer chooses one hypothesis which has small **w**
- We minimize $E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$, this is similar to, $E_{out}(h) \leq E_{in}(h) + \Omega(H)$
- The terms $\frac{\lambda}{N}\Omega(h)$ and $\Omega(H)$ represent the complexity, where $\frac{\lambda}{N}\Omega(h)$ is for individual hypothesis that help to navigate the hypotheses set, and $\Omega(H)$ is the complexity of the hypothesis set
- E_{aug} is better than E_{in} as a proxy for E_{out}

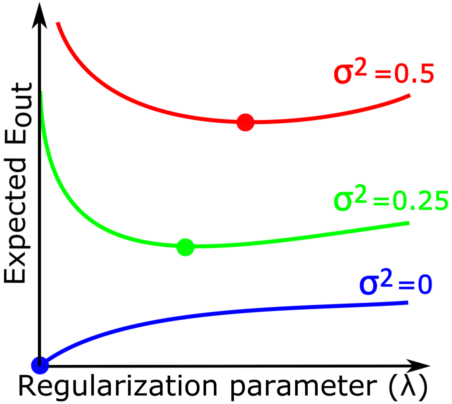
Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

How to choose a regularizer?

- Constraint in the direction of the target function
- Reduce the overfitting through applying a methodology that harms the overfitting more it harms the fitting, i.e. harms the noise than harming the signal
- Move in the direction of smoother or simpler; but, when to stop?
- Use the validation to get the optimal λ

In both stochastic and deterministic noise types:



- With no noise (both types of noise), there is no need for the regularizer
- Increasing the noise, more regularization is needed

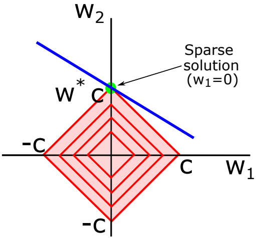
Lecture 12: Regularization

- Review of Lecture 11
- What is the regularization?
- Mathematical and geometrical background
- General form of augmented error
- How to choose a regularizer?
- Common types of regularization

L_1 regularization

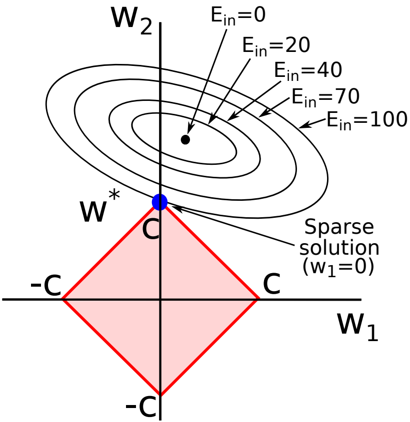
- This term adds the absolute value of the magnitude of coefficients
- This type of regularization can yield *sparse models*, i.e. models with few coefficients because some coefficients become zero and hence eliminated
- Geometrically, this regularizer shrinks some parameters to zero, and hence these eliminated parameters will not play any role in the learning model

- Assume we have a simple problem, $A\mathbf{x} = b$, which is a simple/linear problem. We need two points to fix a line; but, if we have only one point; hence, there are infinite solutions along the line passes through that point
- L_1 for the vector $[x_1, x_2]$ is $|x_1| + |x_2|$ and the L_1 norm equals to a constant c
- To find a solution, the red shape is enlarged by increasing c to touch the solution line. At the touch point, the constant c is the smallest L_1 norm with all possible solutions
- The touch point almost at a vertex of the shape and this is the geometrical interpretation of the sparse solution

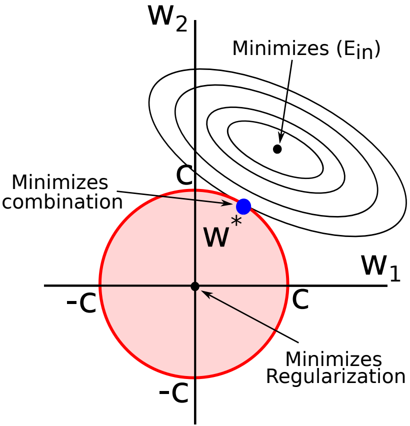


L_2 regularization

- The L_2 regularizer adds penalty equal to the square of the magnitude of coefficients
- In contrast to the L_1 regularizer, L_2 will not yield sparse models and all coefficients are shrunk by the same factor; thus, no coefficients eliminated
- Increasing λ in L_2 reduces the coefficients. This regularizer is used in SVM and Ridge regression
- Assume we have two parameters (w_1 and w_2) in a given problem. In L_1, the constraint functions can be thought of as follows, $|w_1| + |w_2| \leq s$, this implies that the shape L_1 regularizer is represented by a square. The constraint of L_2 can be represented as follows, $w_1^2 + w_2^2 \leq s^2$ and the shape is circle
- L_1 penalizes small parameters more than L_2



(e) L_1 regularizer



(f) L_2 regularizer