ECE595 / STAT598: Machine Learning I
Lecture 31 Regularization

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

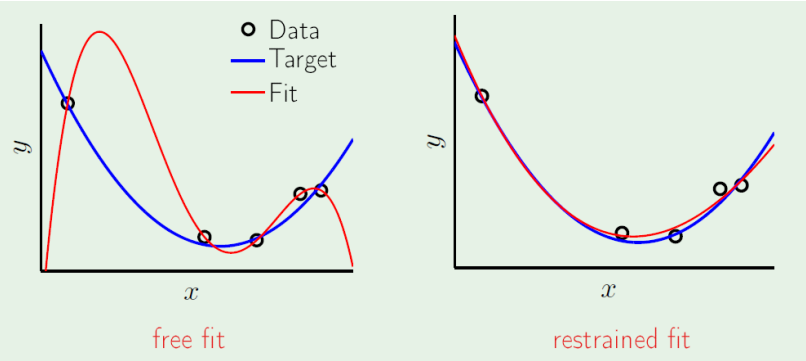
Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

Today's Lecture:

- Motivation for Regularization
  - VC Analysis
  - Example
- Two Regularization Techniques
  - Weight Decay
  - Augmented Error
- Choosing a Regularization
  - Pill or Poisson?
  - Role of λ

Overcoming Overfit



- Regularization is one weapon to combat overfitting.
- Constrains the learning algorithm to improve out-sample error when noise is present.
- Regularization is as much an art as it is a science.

Regularization from VC Analysis

$$E_{\text{out}}(h) \leq E_{\text{in}}(h) + \Omega(\mathcal{H}), \qquad \text{for all } h \in \mathcal{H} \tag{1}$$

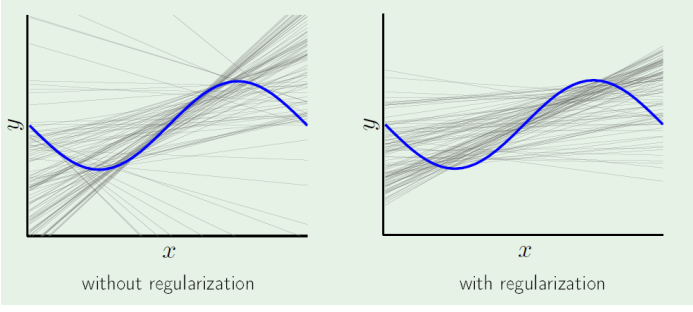
- Model complexity penalty $\Omega(\mathcal{H})$
- If you want $E_{\text{out}}(h)$ to be small, better to make $\Omega(\mathcal{H})$ small
- Roughly speaking, fit data using a "simple" h from $\mathcal{H}$
- So you are effectively minimizing

$$\underset{h}{\text{minimize}} \quad E_{\text{in}}(h) + \Omega(h),$$

- That is, instead of minimizing $E_{\text{in}}(h)$ only, you minimize $\Omega(h)$ too

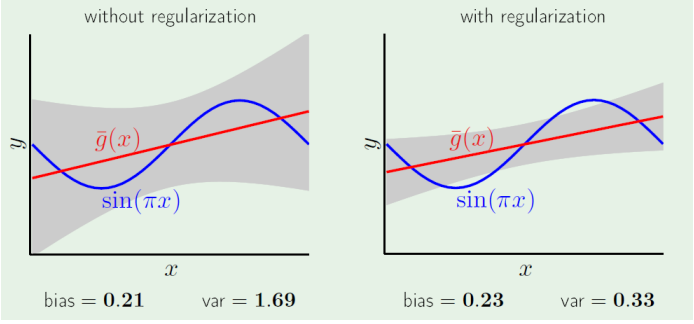
Example

- One regularization technique is **weight decay**.
- Measures the complexity of a hypothesis h by the size of the coefficients used to represent h (e.g., in a linear model).
- This technique prefers mild lines with small offset and slope.
- Applying this concept to the sine example before, trying to fit $N = 2$ data points, using $\mathcal{H}_1$ (the set of lines).



Example

- Recall the constant model: Fit the two data points using a constant line.
- Constant model has $E_{\text{out}} = 0.75$
- Unregularized model has $E_{\text{out}} = 1.90$
- Regularized model has $E_{\text{out}} = 0.56$
- Bias-variance: Improve variance but suffer from bias. Overall is better.



Why Need Regularization?

- The linear model is too sophisticated for the amount of data we have.
- A line can fit any two points!
- This problem is still here even if we change the target function.
- The need of regularization depends on **quantity** of data, and **quality** of data.
- Given only two points, we can either choose
  - a simple model, e.g., constant model
  - to constrain the model, e.g., weight decay
- Constraining the model gives us more flexibility.

Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

Today's Lecture:

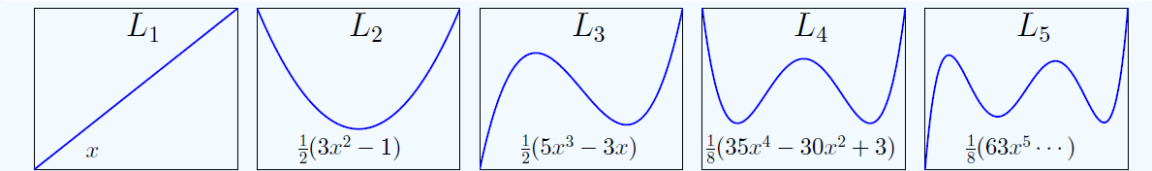
- Motivation for Regularization
  - VC Analysis
  - Example
- Two Regularization Techniques
  - Weight Decay
  - Augmented Error
- Choosing a Regularization
  - Pill or Poisson?
  - Role of λ

Soft Order Constraint

Consider the following example

- $\mathcal{H}$ = set of polynomials in one variable $x \in [-1, 1]$.
- E.g., $h(x) = 2x^2 + 3x + 7$.
- Want to express $h(x)$ using basis function.
- Basis functions for polynomials are Legendre polynomials $L_q(x)$, $q = 1, 2, \ldots$
- So, any $h(x)$ can be expressed as

$$h(x) = \sum_{q=1}^{Q} w_q L_q(x) \tag{2}$$



Soft Order Constraint

This model is indeed linear! (Why?)

- You define a nonlinear transform Φ,

$$\boldsymbol{z} = \Phi(x) = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix}$$

- The hypothesis set is

$$\mathcal{H}_Q = \left\{ h \middle| h(x) = \boldsymbol{w}^T \boldsymbol{z} = \sum_{q=0}^{Q} w_q L_q(x) \right\}$$

- So now you can define training error (for linear regression) as

$$E_{\text{in}}(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{w}^T \boldsymbol{z}_n - y_n)^2$$

Soft Order Constraint

There are multiple ways of constraining the weights.

- **Hard** constraint:
  - Force coefficients to be zero.
  - For example,
$$\mathcal{H}_2 = \{\boldsymbol{w} \mid \boldsymbol{w} \in \mathcal{H}_{10}; w_q = 0, \text{for} q \geq 3\}.$$
- **Soft** constraint:
  - Force coefficients to be small.
  - For example,
$$\sum_{q=0}^{Q} w_q^2 \leq C$$
  - It encourages weights to be small without changing the order of the polynomial by explicitly forcing some weights to zero.

VC Perspective of Soft Order Constraint

- The optimization is

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad E_{\text{in}}(\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{w}^T\boldsymbol{w} \leq C \tag{3}$$

- We know $E_{\text{in}}(\boldsymbol{w}) = \frac{1}{N}\|\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{y}\|^2$
- The hypothesis set is

$$\mathcal{H}(C) = \{h \mid h(x) = \boldsymbol{w}^T\boldsymbol{z},\ \boldsymbol{w}^T\boldsymbol{w} \leq C\}$$

- So the optimization is equivalent to minimize E_{in} over $\mathcal{H}(C)$
- If $C_1 < C_2$, then $\mathcal{H}(C_1) \subset \mathcal{H}(C_2)$ and $d_{\text{vc}}(\mathcal{H}(C_1)) \leq d_{\text{vc}}(\mathcal{H}(C_2))$
- So we should expect better generalization with $\mathcal{H}(C_1)$

Solving the Soft Order Constraint Problem

The optimization problem is

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \frac{1}{N}\|\boldsymbol{Z}\boldsymbol{w}-\boldsymbol{y}\|^2 \quad \text{subject to} \quad \boldsymbol{w}^T\boldsymbol{w} \leq C \tag{4}$$

- Using Lagrangian techniques we can show that the minimization is equivalent to

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad E_{\text{in}}(\boldsymbol{w}) + \frac{\lambda_C}{N}\boldsymbol{w}^T\boldsymbol{w}$$

  for some choices of λ_C.
- You can further change the constraint to

$$\sum_{q=0}^{Q} \gamma_q w_q^2 \leq C$$

  - $\gamma_q = q$ or $\gamma_q = e^q$ encourages a low-order fit
  - $\gamma_q = (1+q)^{-1}$ or $\gamma_q = e^{-q}$ encourages a high-order fit

Augmented Error

Another type of regularization is **augmented error**

$$E_{\text{aug}}(\boldsymbol{w}) = E_{\text{in}}(\boldsymbol{w}) + \lambda \boldsymbol{w}^T \boldsymbol{w} \tag{5}$$

- Unconstrained minimization is often easier than constrained minimization
- But you are paying the price of interpretability
- For a given C, soft order constraint corresponds to selecting a hypothesis from a smaller set $\mathcal{H}(C)$
- VC analysis says we will get a better generalization when C decreases (but not too much)
- The optimal C is sum square magnitude we allow.
- For augmented error, you need to find the optimal parameter λ^*
- This is not very interpretable.

VC Perspective of Augmented Error

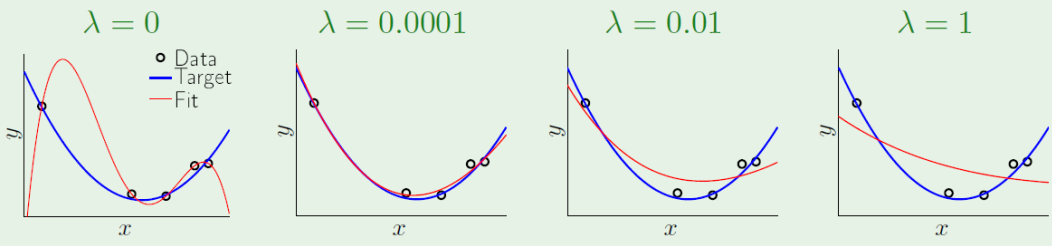
The augmented error for a hypothesis $h \in \mathcal{H}$ is

$$E_{\text{aug}}(h, \lambda, \Omega) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h) \tag{6}$$

- Here, $\Omega(h) = \boldsymbol{w}^T\boldsymbol{w}$
- There are two components of the penalty:
  - The regularizer $\Omega(h)$ which penalizes a particular property of h
  - The regularization parameter λ which controls the amount of regularization
- As N increases, the need for regularization goes down
- This equation resembles VC bound

Choice of λ

Minimizing $E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$ for different λ's:



Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

Today's Lecture:

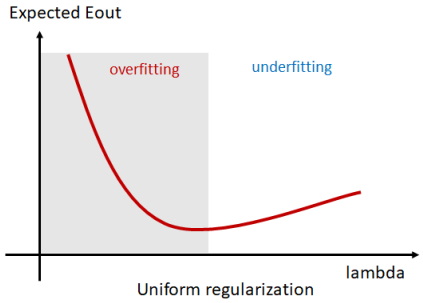
- Motivation for Regularization
  - VC Analysis
  - Example
- Two Regularization Techniques
  - Weight Decay
  - Augmented Error
- Choosing a Regularization
  - Pill or Poisson?
  - Role of λ

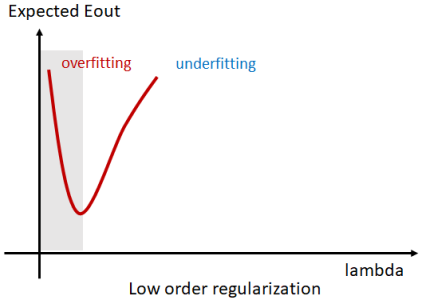
Choosing a Regularization: Pill or Poisson?

- Regularization = choose $\Omega(h)$ and λ.
- Choice of $\Omega(h)$ is heuristic.
- Finding a perfect Ω is as difficult as finding a perfect $\mathcal{H}$.
- Some forms of regularization work and some do not.
- Too little: Underfitting. Too much: Overfitting/
- Why bother with regularization if so many choices can go wrong?
- Regularization is a **necessary** evil.
- If our model is too sophisticated for the amount of data we have, we are doomed.
- By applying regularization, we have a chance.

Overfit and Underfit

- Consider a 15-th order polynomial. So $\mathcal{H}_{15}$.
- Two choices of regularization:
  - Uniform regularization: $\Omega_{\text{uniform}}(\boldsymbol{w}) = \sum_{q=0}^{15} w_q^2$
  - Low-order regularization: $\Omega_{\text{low}}(\boldsymbol{w}) = \sum_{q=0}^{15} q w_q^2$
- When λ too small, overfit. When λ too large, underfit.
- For optimal λ, the two are quite similar.

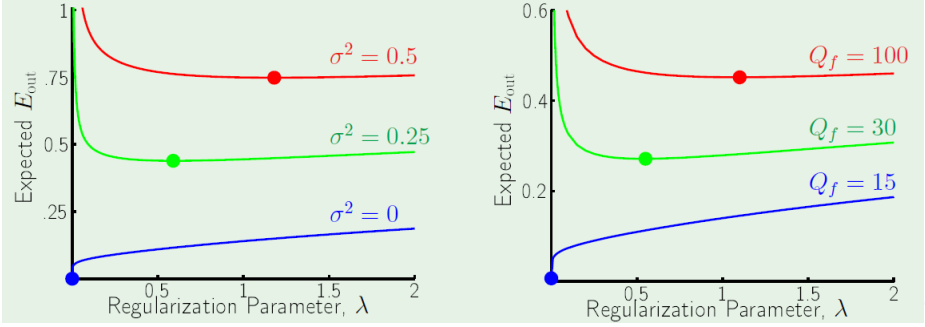




Regularization on Noise and Target Complexity

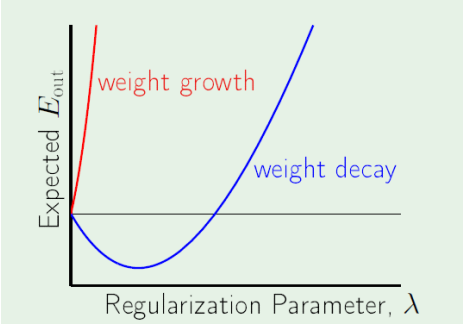
Let us analyze the impact of regularization to noise

- Noise: Uncertainty in each measured data. Measured in terms of σ^2.
- If you have noise, then you need to adjust λ depending on the noise level.
- Target complexity: Suppose data comes from $\mathcal{H}_{15}$ but you use $\mathcal{H}_{50}$. Measured in terms of Q_f.
- Like noise, you need to adjust λ to optimize generalization.



What if Picked a Wrong Regularization?

- Suppose we should encourage low-order coefficients, but the regularization promotes high-order coefficients.
- Are we screwed?
- No, you still have the regularization parameter λ.
- Below is an example.
- Choosing the regularization parameter can be done using validation. Will discuss next.



Summary

- Whenever you train a model, try including regularization.
- It can be as simple as $\boldsymbol{w}^T\boldsymbol{w}$.
- Helps dramatically when there is noise in data, not enough data, complex target.
- Hand-waving argument: noise is high frequency. Complex target is also high frequency.
- So low-frequency regularization helps.
- As long as you have a good λ, the benefit of regularization is often more than the harm.
- Modern deep learning can easily incorporate regularization.
- E.g., you can regularize the magnitude of the network weights, or number of non-zeros through sparsity.

Reading List

- Yaser Abu-Mostafa, Learning from Data, chapter 4.2
- Stanford CS 229 http://cs229.stanford.edu/notes/cs229-notes5.pdf