

Machine Learning Foundations

(機器學習基石)



Lecture 14: Regularization

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)



Roadmap

- 1 When Can Machines Learn?
- 2 Why Can Machines Learn?
- 3 How Can Machines Learn?
- 4 How Can Machines Learn **Better**?

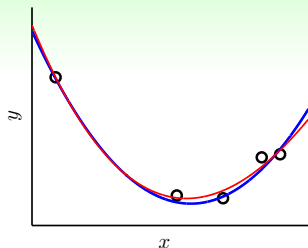
Lecture 13: Hazard of Overfitting

overfitting happens with **excessive power**, **stochastic/deterministic noise**, and **limited data**

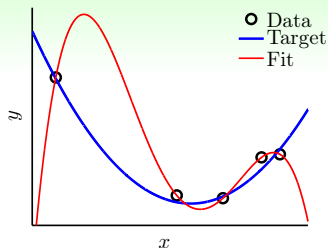
Lecture 14: Regularization

- Regularized Hypothesis Set
- Weight Decay Regularization
- Regularization and VC Theory
- General Regularizers

Regularization: The Magic

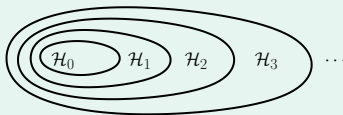


'regularized fit'



overfit

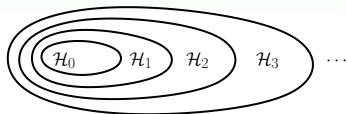
- idea: 'step back' from \mathcal{H}_{10} to \mathcal{H}_2



- name history: function approximation for **ill-posed problems**

how to step back?

Stepping Back as Constraint



Q -th order polynomial **transform** for $x \in \mathbb{R}$:

$$\Phi_Q(x) = (1, x, x^2, \dots, x^Q)$$

+ **linear regression**, denote $\tilde{\mathbf{w}}$ by \mathbf{w}

hypothesis \mathbf{w} in \mathcal{H}_{10} : $w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \dots + w_{10} x^{10}$

hypothesis \mathbf{w} in \mathcal{H}_2 : $w_0 + w_1 x + w_2 x^2$

that is, $\mathcal{H}_2 = \mathcal{H}_{10}$ AND 'constraint that $w_3 = w_4 = \dots = w_{10} = 0$ '

step back = **constraint**

Regression with Constraint

$$\mathcal{H}_{10} \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right\}$$

regression with \mathcal{H}_{10} :

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$

$$\mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right. \\ \left. \text{while } w_3 = w_4 = \dots = w_{10} = 0 \right\}$$

regression with \mathcal{H}_2 :

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \\ \text{s.t.} \quad w_3 = w_4 = \dots = w_{10} = 0 \end{aligned}$$

step back = constrained optimization of E_{in}

why don't you just use $\mathbf{w} \in \mathbb{R}^{2+1}$? :-)

Regression with Looser Constraint

$$\mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right. \\ \left. \text{while } w_3 = \dots = w_{10} = 0 \right\}$$

regression with \mathcal{H}_2 :

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$

$$\text{s.t. } w_3 = \dots = w_{10} = 0$$

$$\mathcal{H}'_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right. \\ \left. \text{while } \geq 8 \text{ of } w_q = 0 \right\}$$

regression with \mathcal{H}'_2 :

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$

$$\text{s.t. } \sum_{q=0}^{10} \mathbb{I}[w_q \neq 0] \leq 3$$

- more flexible than \mathcal{H}_2 : $\mathcal{H}_2 \subset \mathcal{H}'_2$
- less risky than \mathcal{H}_{10} : $\mathcal{H}'_2 \subset \mathcal{H}_{10}$

bad news for sparse hypothesis set \mathcal{H}'_2 :
NP-hard to solve :-()

Regression with Softer Constraint

$$\mathcal{H}'_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right. \\ \left. \text{while } \geq 8 \text{ of } w_q = 0 \right\}$$

regression with \mathcal{H}'_2 :

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} \mathbb{I}[w_q \neq 0] \leq 3$$

$$\mathcal{H}(C) \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right. \\ \left. \text{while } \|\mathbf{w}\|^2 \leq C \right\}$$

regression with $\mathcal{H}(C)$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} w_q^2 \leq C$$

- $\mathcal{H}(C)$: overlaps but not exactly the same as \mathcal{H}'_2
- soft and smooth structure over $C \geq 0$:
 $\mathcal{H}(0) \subset \mathcal{H}(1.126) \subset \dots \subset \mathcal{H}(1126) \subset \dots \subset \mathcal{H}(\infty) = \mathcal{H}_{10}$

regularized hypothesis \mathbf{w}_{REG} :
 optimal solution from
 regularized hypothesis set $\mathcal{H}(C)$

Fun Time

For $Q \geq 1$, which of the following hypothesis (weight vector $\mathbf{w} \in \mathbb{R}^{Q+1}$) is not in the regularized hypothesis set $\mathcal{H}(1)$?

① $\mathbf{w}^T = [0, 0, \dots, 0]$

② $\mathbf{w}^T = [1, 0, \dots, 0]$

③ $\mathbf{w}^T = [1, 1, \dots, 1]$

④ $\mathbf{w}^T = \left[\sqrt{\frac{1}{Q+1}}, \sqrt{\frac{1}{Q+1}}, \dots, \sqrt{\frac{1}{Q+1}} \right]$

Matrix Form of Regularized Regression Problem

$$\begin{aligned}
 \min_{\mathbf{w} \in \mathbb{R}^{Q+1}} \quad & E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{z}_n - y_n)^2 \\
 & \underbrace{\hspace{10em}}_{(\mathbf{Z}\mathbf{w} - \mathbf{y})^T (\mathbf{Z}\mathbf{w} - \mathbf{y})} \\
 \text{s.t.} \quad & \underbrace{\sum_{q=0}^Q w_q^2}_{\mathbf{w}^T \mathbf{w}} \leq C
 \end{aligned}$$

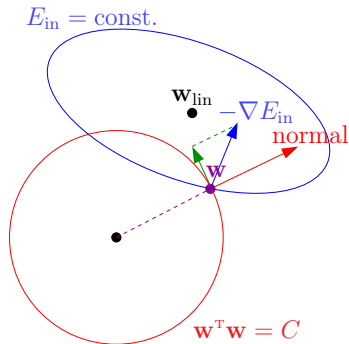
- $\sum_n \dots = (\mathbf{Z}\mathbf{w} - \mathbf{y})^T (\mathbf{Z}\mathbf{w} - \mathbf{y})$, **remember? :-)**
- $\mathbf{w}^T \mathbf{w} \leq C$: feasible \mathbf{w} within a radius- \sqrt{C} hypersphere

how to solve
constrained optimization problem?

The Lagrange Multiplier

$$\min_{\mathbf{w} \in \mathbb{R}^{Q+1}} E_{\text{in}}(\mathbf{w}) = \frac{1}{N}(\mathbf{Z}\mathbf{w} - \mathbf{y})^T(\mathbf{Z}\mathbf{w} - \mathbf{y}) \text{ s.t. } \mathbf{w}^T \mathbf{w} \leq C$$

- decreasing direction: $-\nabla E_{\text{in}}(\mathbf{w})$, **remember? :-)**
- normal** vector of $\mathbf{w}^T \mathbf{w} = C$: \mathbf{w}
- if $-\nabla E_{\text{in}}(\mathbf{w})$ and \mathbf{w} not parallel: can **decrease** $E_{\text{in}}(\mathbf{w})$ **without violating the constraint**
- at optimal solution \mathbf{w}_{REG} ,
 $-\nabla E_{\text{in}}(\mathbf{w}_{\text{REG}}) \propto \mathbf{w}_{\text{REG}}$



want: find **Lagrange multiplier** $\lambda > 0$ and \mathbf{w}_{REG}
 such that $\nabla E_{\text{in}}(\mathbf{w}_{\text{REG}}) + \frac{2\lambda}{N} \mathbf{w}_{\text{REG}} = \mathbf{0}$

Augmented Error

- if **oracle** tells you $\lambda > 0$, then

solving
$$\nabla E_{\text{in}}(\mathbf{w}_{\text{REG}}) + \frac{2\lambda}{N} \boxed{\mathbf{w}_{\text{REG}}} = \mathbf{0}$$

$$\frac{2}{N} (Z^T Z \mathbf{w}_{\text{REG}} - Z^T \mathbf{y}) + \frac{2\lambda}{N} \boxed{\mathbf{w}_{\text{REG}}} = \mathbf{0}$$

- optimal solution:

$$\mathbf{w}_{\text{REG}} \leftarrow (Z^T Z + \lambda \mathbf{I})^{-1} Z^T \mathbf{y}$$

—called **ridge regression** in Statistics

minimizing **unconstrained** E_{aug} effectively
minimizes some **C-constrained** E_{in}

Augmented Error

- if **oracle** tells you $\lambda > 0$, then

solving
$$\nabla E_{\text{in}}(\mathbf{w}_{\text{REG}}) + \frac{2\lambda}{N} \boxed{\mathbf{w}_{\text{REG}}} = \mathbf{0}$$

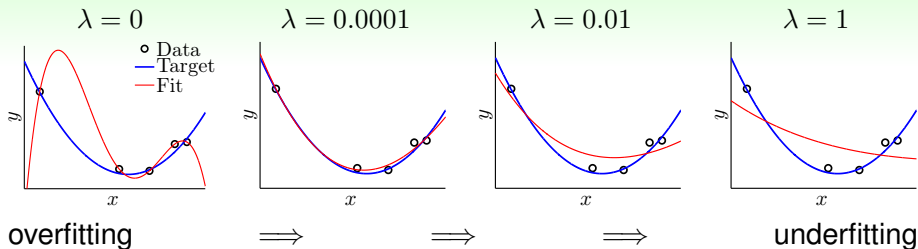
equivalent to minimizing
$$\underbrace{E_{\text{in}}(\mathbf{w})}_{\text{augmented error } E_{\text{aug}}(\mathbf{w})} + \frac{\lambda}{N} \overbrace{\mathbf{w}^T \mathbf{w}}^{\text{regularizer}}$$

- regularization with **augmented error** instead of **constrained** E_{in}

$$\mathbf{w}_{\text{REG}} \leftarrow \underset{\mathbf{w}}{\text{argmin}} E_{\text{aug}}(\mathbf{w}) \text{ for given } \lambda > 0 \text{ or } \lambda = 0$$

minimizing **unconstrained** E_{aug} effectively
minimizes some **C-constrained** E_{in}

The Results



philosophy: *a little regularization goes a long way!*

call ' $+\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$ ' **weight-decay** regularization:

larger λ

\iff prefer shorter \mathbf{w}

\iff effectively smaller C

—go with 'any' transform + linear model

Some Detail: Legendre Polynomials

$$\min_{\mathbf{w} \in \mathbb{R}^{Q+1}} \frac{1}{N} \sum_{n=0}^N (\mathbf{w}^T \boldsymbol{\Phi}(x_n) - y_n)^2 + \frac{\lambda}{N} \sum_{q=0}^Q w_q^2$$

naïve polynomial **transform**:

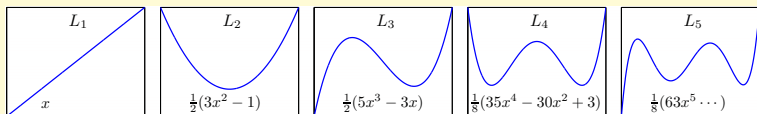
$$\boldsymbol{\Phi}(\mathbf{x}) = (1, x, x^2, \dots, x^Q)$$

—when $x_n \in [-1, +1]$, x_n^q really small, needing large w_q

normalized polynomial **transform**:

$$(1, L_1(x), L_2(x), \dots, L_Q(x))$$

—‘orthonormal basis functions’
called **Legendre polynomials**



Fun Time

When would \mathbf{w}_{REG} equal \mathbf{w}_{LIN} ?

- 1 $\lambda = 0$
- 2 $C = \infty$
- 3 $C \geq \|\mathbf{w}_{\text{LIN}}\|^2$
- 4 all of the above

Regularization and VC Theory

Regularization by
Constrained-Minimizing E_{in}

$$\min_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \mathbf{w}^T \mathbf{w} \leq C$$



VC Guarantee of
Constrained-Minimizing E_{in}

$$E_{\text{out}}(\mathbf{w}) \leq E_{\text{in}}(\mathbf{w}) + \Omega(\mathcal{H}(C))$$



C equivalent to some λ

Regularization by
Minimizing E_{aug}

$$\min_{\mathbf{w}} E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$$

minimizing E_{aug} : indirectly getting VC
guarantee **without confining to $\mathcal{H}(C)$**

Another View of Augmented Error

Augmented Error

$$E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$$

VC Bound

$$E_{\text{out}}(\mathbf{w}) \leq E_{\text{in}}(\mathbf{w}) + \Omega(\mathcal{H})$$

- regularizer $\mathbf{w}^T \mathbf{w}$: complexity of a single hypothesis
- generalization price $\Omega(\mathcal{H})$: complexity of a hypothesis set
- if $\frac{\lambda}{N} \Omega(\mathbf{w})$ 'represents' $\Omega(\mathcal{H})$ well,
 E_{aug} is a better proxy of E_{out} than E_{in}

minimizing E_{aug} :

(heuristically) operating with the better proxy;
(technically) enjoying flexibility of whole \mathcal{H}

Effective VC Dimension

$$\min_{\mathbf{w} \in \mathbb{R}^{\tilde{d}+1}} E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \Omega(\mathbf{w})$$

- model complexity?

$d_{\text{VC}}(\mathcal{H}) = \tilde{d} + 1$, because $\{\mathbf{w}\}$ ‘**all considered**’ during minimization

- $\{\mathbf{w}\}$ ‘**actually needed**’: $\mathcal{H}(\mathcal{C})$, with some \mathcal{C} equivalent to λ

- $d_{\text{VC}}(\mathcal{H}(\mathcal{C}))$:

effective VC dimension $d_{\text{EFF}}(\mathcal{H}, \underbrace{\mathcal{A}}_{\min E_{\text{aug}}})$

explanation of regularization:

$d_{\text{VC}}(\mathcal{H})$ large,

while $d_{\text{EFF}}(\mathcal{H}, \mathcal{A})$ small if \mathcal{A} regularized

Fun Time

Consider the weight-decay regularization with regression. When increasing λ in \mathcal{A} , what would happen with $d_{\text{EFF}}(\mathcal{H}, \mathcal{A})$?

- ① $d_{\text{EFF}} \uparrow$
- ② $d_{\text{EFF}} \downarrow$
- ③ $d_{\text{EFF}} = d_{\text{VC}}(\mathcal{H})$ and does not depend on λ
- ④ $d_{\text{EFF}} = 1126$ and does not depend on λ

General Regularizers $\Omega(\mathbf{w})$

want: constraint in the **'direction' of target function**

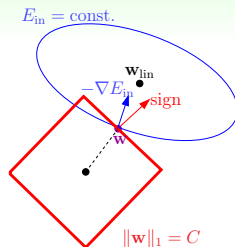
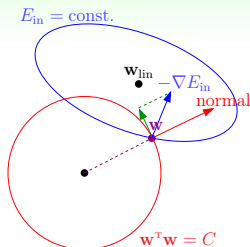
- target-dependent: some **properties** of target, if known
 - symmetry** regularizer: $\sum \llbracket q \text{ is odd} \rrbracket w_q^2$
- plausible: direction towards **smoother** or **simpler**
stochastic/deterministic noise both **non-smooth**
 - sparsity** (L1) regularizer: $\sum |w_q|$ (next slide)
- friendly: easy to **optimize**
 - weight-decay** (L2) regularizer: $\sum w_q^2$
- bad? :-)**: no worries, guard by λ

augmented error = error $\widehat{\text{err}}$ + regularizer Ω

regularizer: **target-dependent**, **plausible**, or **friendly**
ringing a bell? :-)

error measure: **user-dependent**, **plausible**, or **friendly**

L2 and L1 Regularizer



L2 Regularizer

$$\Omega(\mathbf{w}) = \sum_{q=0}^Q w_q^2 = \|\mathbf{w}\|_2^2$$

- convex, differentiable everywhere
- easy to optimize

L1 Regularizer

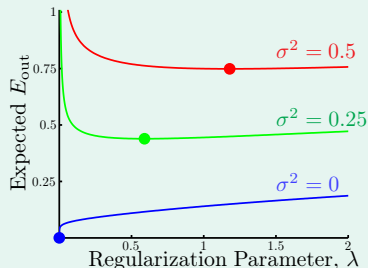
$$\Omega(\mathbf{w}) = \sum_{q=0}^Q |w_q| = \|\mathbf{w}\|_1$$

- convex, **not** differentiable everywhere
- **sparsity** in solution

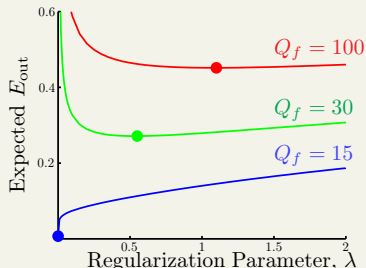
L1 useful if needing **sparse solution**

The Optimal λ

stochastic noise



deterministic noise



- more noise \iff more regularization needed
—more bumpy road \iff putting brakes more
- noise **unknown**—important to **make proper choices**

how to choose?

stay tuned for the next lecture! :-)

Fun Time

Consider using a regularizer $\Omega(\mathbf{w}) = \sum_{q=0}^Q 2^q w_q^2$ to work with Legendre polynomial regression. Which kind of hypothesis does the regularizer prefer?

- ① symmetric polynomials satisfying $h(x) = h(-x)$
- ② low-dimensional polynomials
- ③ high-dimensional polynomials
- ④ no specific preference

Summary

- 1 When Can Machines Learn?
- 2 Why Can Machines Learn?
- 3 How Can Machines Learn?
- 4 How Can Machines Learn **Better**?

Lecture 13: Hazard of Overfitting

Lecture 14: Regularization

- Regularized Hypothesis Set
original \mathcal{H} + constraint
- Weight Decay Regularization
add $\frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$ in E_{aug}
- Regularization and VC Theory
regularization decreases d_{EFF}
- General Regularizers
target-dependent, [plausible], or [friendly]

- **next: choosing from the so-many models/parameters**