Chapter ML:III (continued)

III. Linear Models

- ❑ Logistic Regression
- ❑ Loss Computation in Detail
- ❑ Overfitting
- ❑ Regularization
- ❑ Gradient Descent in Detail

Overfitting

Definition 9 (Overfitting)

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h on D.

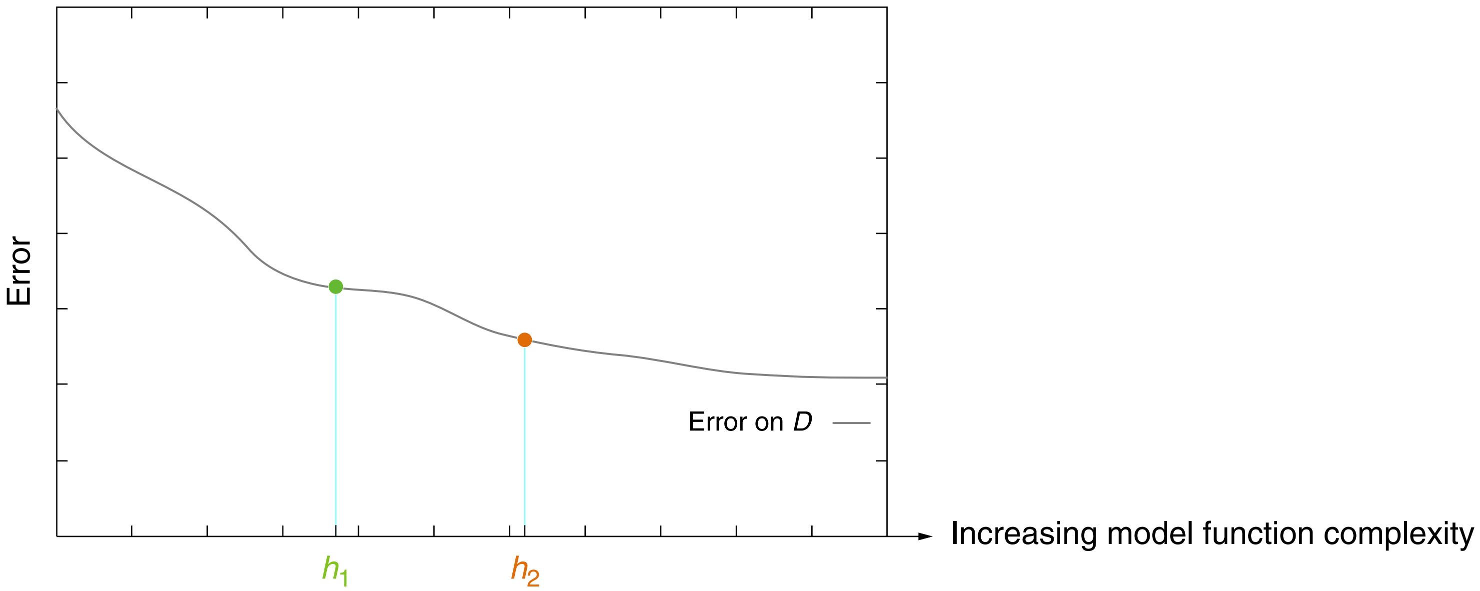
Overfitting

Definition 9 (Overfitting)

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h on D.



Overfitting

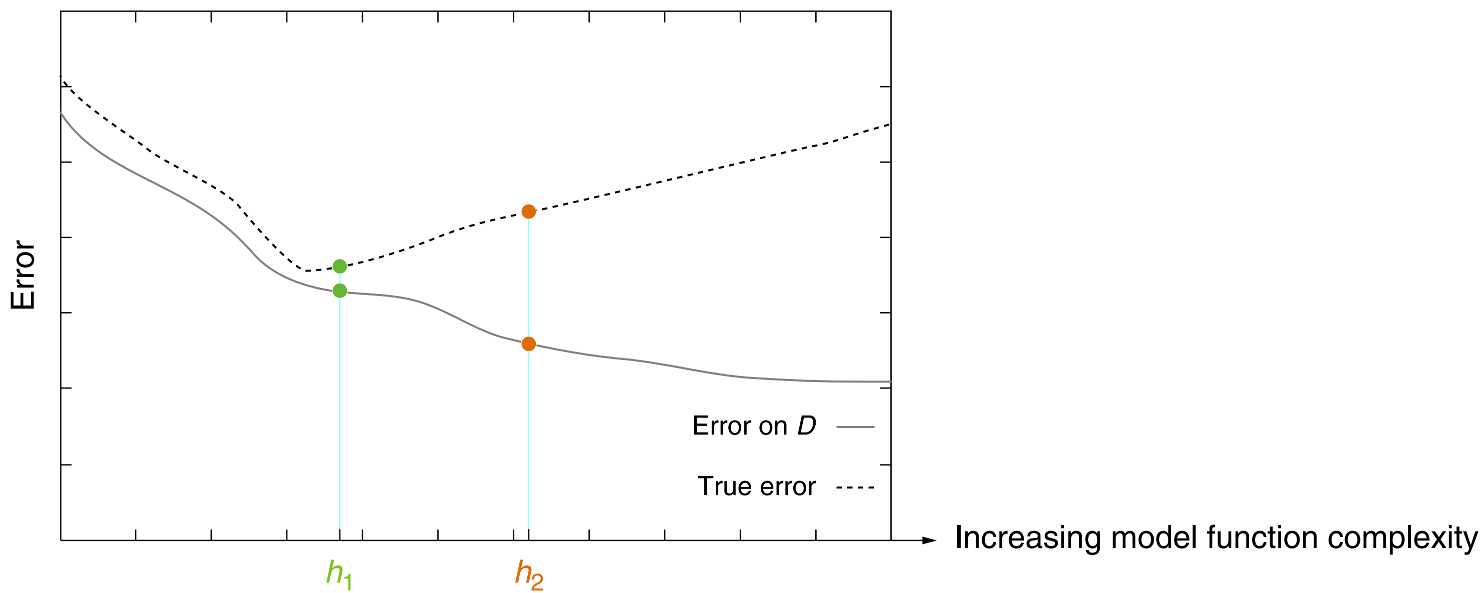
Definition 9 (Overfitting)

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h on D.

[see continuation]



Overfitting

Definition 9 (Overfitting)

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h on D.

[see continuation]

Reasons for overfitting are often rooted in the example set D :

- D is noisy and we "learn noise."
- D is biased and hence not representative.
- D is too small and hence pretends unrealistic data properties.

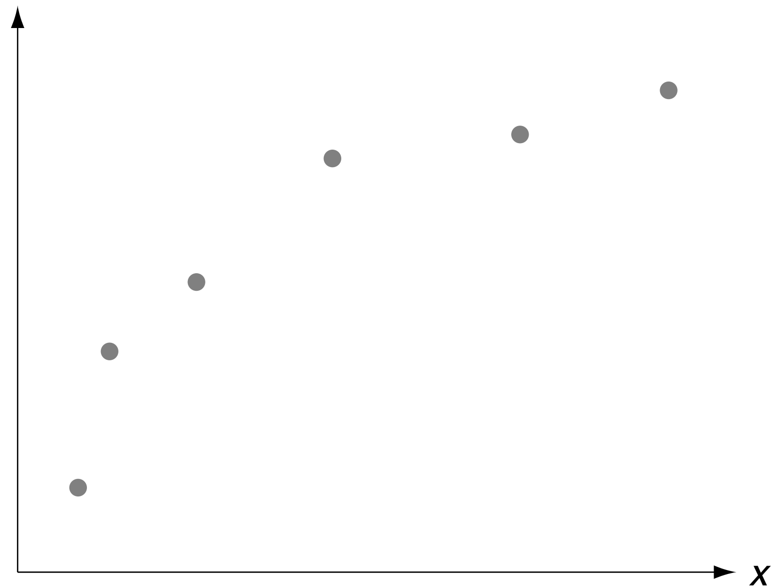
Remarks:

- Recap. A hypothesis is a proposed explanation for a phenomenon. [Wikipedia]

  Here, a hypothesis “explains” (= fits) the data D. Hence, a concrete model function $y()$, $\mathbf{y}()$, or, if the function type is clear from the context, its parameters $\mathbf{w}$ or $\boldsymbol{\theta}$ are called “hypothesis”. The variable name h (similarly: h_1, h_2, h_i, h', etc.) may be used to refer to a specific instance of a model function or its parameters.

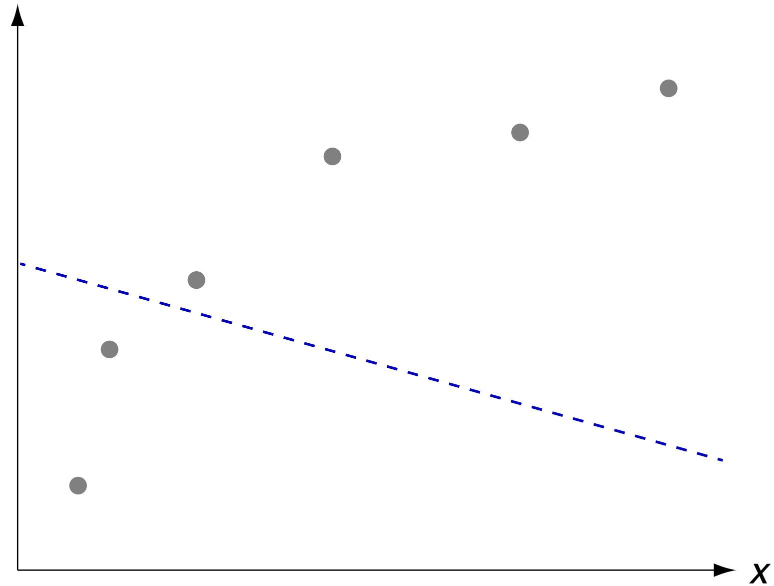
Overfitting

Example: Linear Regression



Overfitting

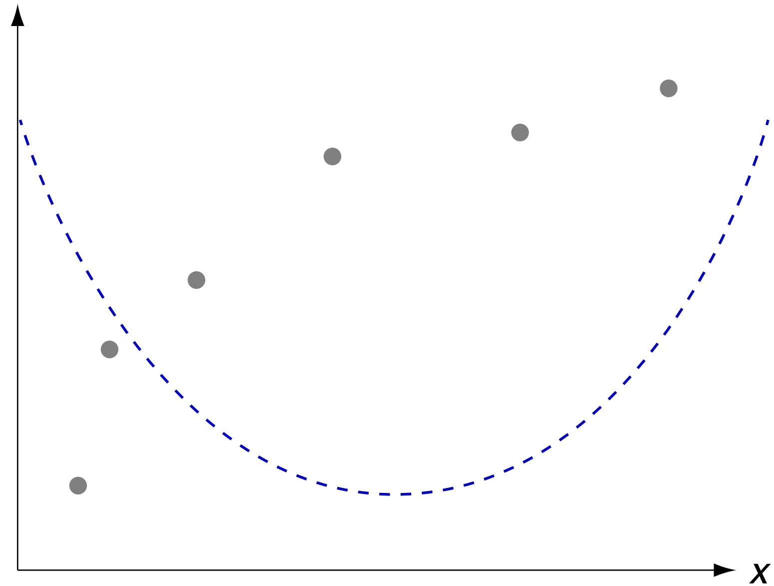
Example: Linear Regression (continued)



(a) $y(x) = w_0 + w_1 \cdot x$

Overfitting

Example: Linear Regression (continued)



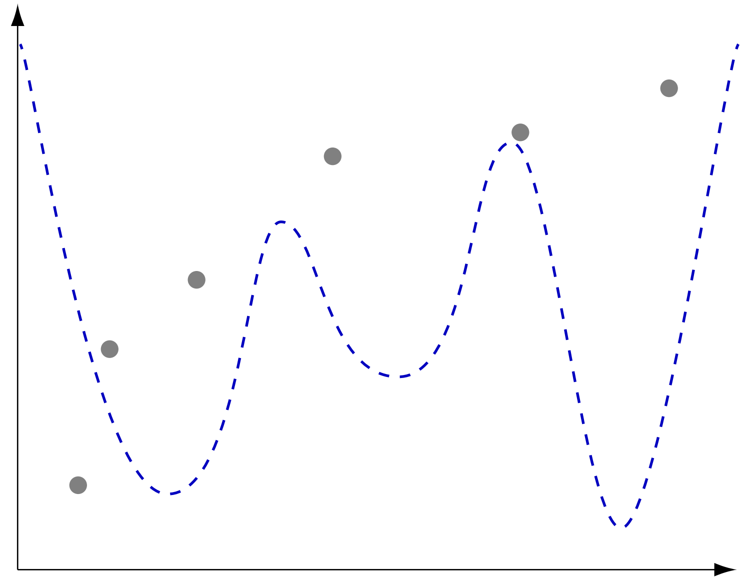
(b) $y(x) = w_0 + w_1 \cdot x + w_2 \cdot x^2$ (basis expansion)

$$y(x) \;=\; (w_0 \;\; w_1 \;\; w_2)\begin{pmatrix}1\\ x\\ x^2\end{pmatrix} \;=:\; \mathbf{w}^T\begin{pmatrix}x_0\\ x_1\\ x_2\end{pmatrix} \;=\; \mathbf{w}^T\mathbf{x} \;=\; y(\mathbf{x}),$$

where $x_0 = 1,\; x_1 = x,\; x_2 = x^2$

Overfitting

Example: Linear Regression (continued)



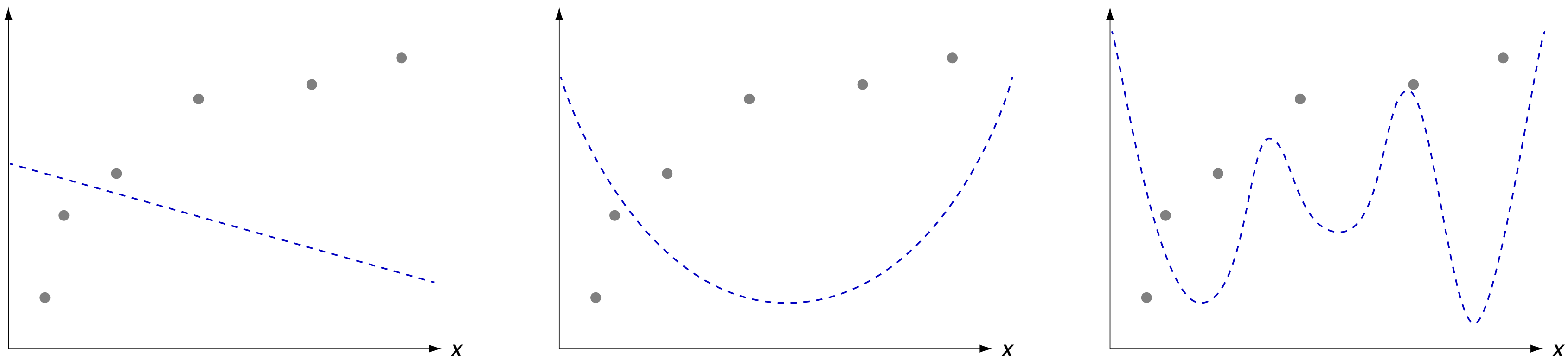
(c) $$y(x) = w_0 + \sum_{j=1}^{6} w_j \cdot x^j$$ (basis expansion)

$y(x) \;=:\; \mathbf{w}^T\mathbf{x} \;=\; y(\mathbf{x}),$ where $x_0 = 1,\; x_j = x^j,\; j = 1, \ldots, 6$

Overfitting

Example: Linear Regression (continued)

Given the three polynomial model functions of degrees 1, 2, and 6, and a training set D_{tr}, select the function that best fits the data:

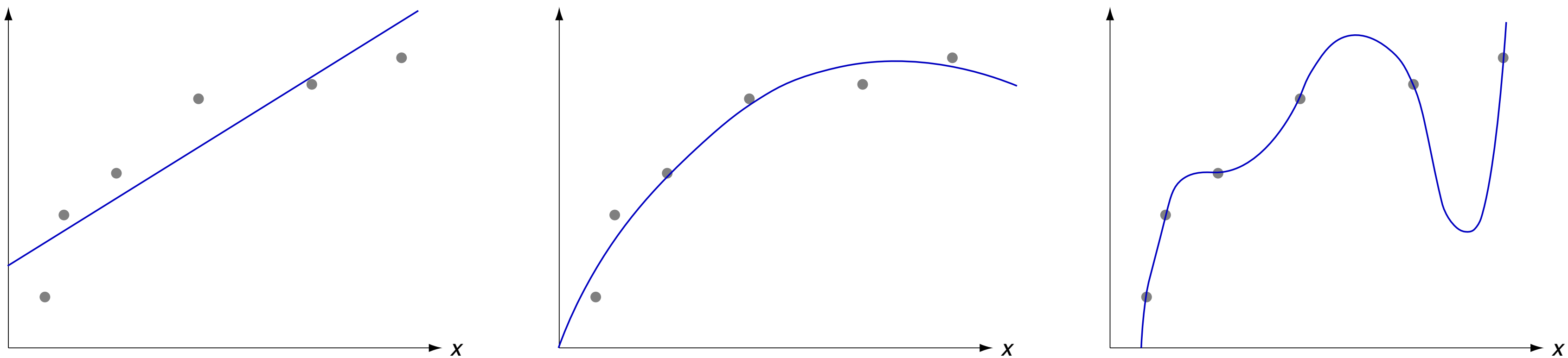


(a) (b) (c)

Overfitting

Example: Linear Regression (continued)

Given the three polynomial model functions of degrees 1, 2, and 6, and a training set D_{tr}, select the function that best fits the data:

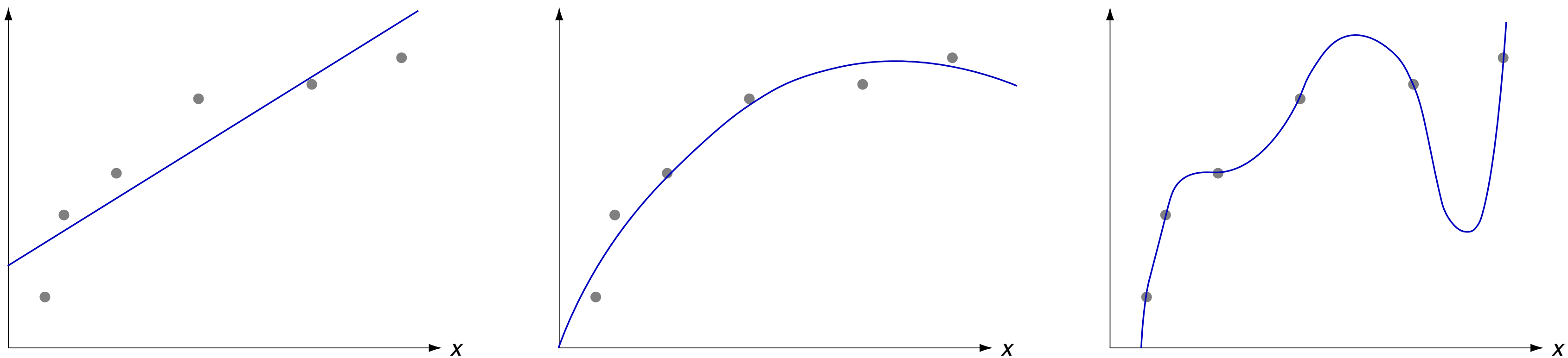


(a) RSS$(\mathbf{w}) \gg 0$ (b) RSS$(\mathbf{w}) > 0$ (c) RSS$(\mathbf{w}) = 0$

Overfitting

Example: Linear Regression (continued)

Given the three polynomial model functions of degrees 1, 2, and 6, and a training set D_{tr}, select the function that best fits the data:



(a) RSS$(\mathbf{w}) \gg 0$ (b) RSS$(\mathbf{w}) > 0$ (c) RSS$(\mathbf{w}) = 0$

Questions:

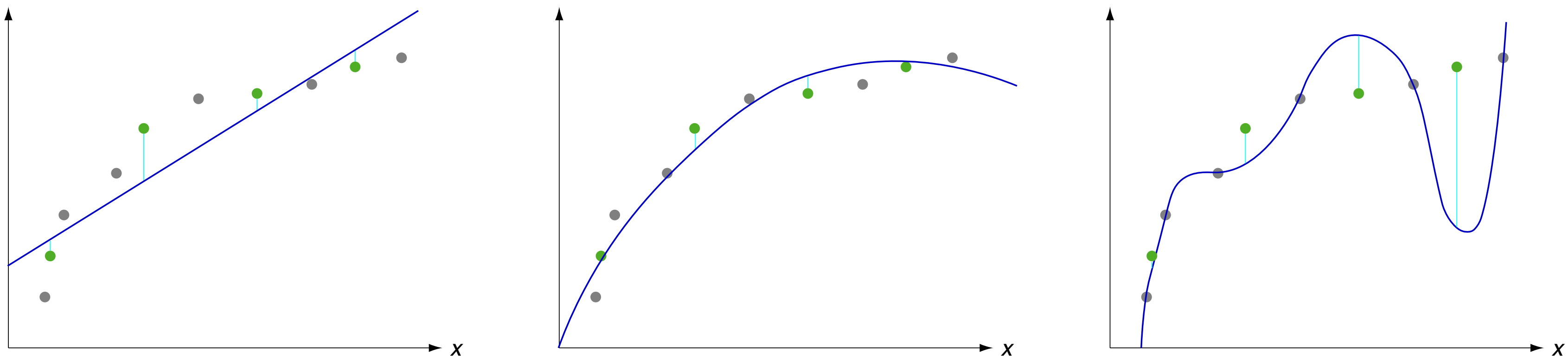
(1) How to choose a suited model function / hypothesis space H?

(2) How to parameterize a model function / pick an element from H?

Overfitting

Example: Linear Regression (continued)

Given the three polynomial model functions of degrees 1, 2, and 6, and a training set D_{tr}, select the function that best fits the data:



(a) $\text{RSS}(\mathbf{w}) \gg 0$ (b) $\text{RSS}(\mathbf{w}) > 0$ (c) $\text{RSS}(\mathbf{w}) \gg 0$

Let D_{test} be a set of test examples.

If $D = D_{tr} \cup D_{test}$ is representative of the real-world population in X, the quadratic model function (b), $y(x) = w_0 + w_1 \cdot x + w_2 \cdot x^2$, is the closest match.

Overfitting

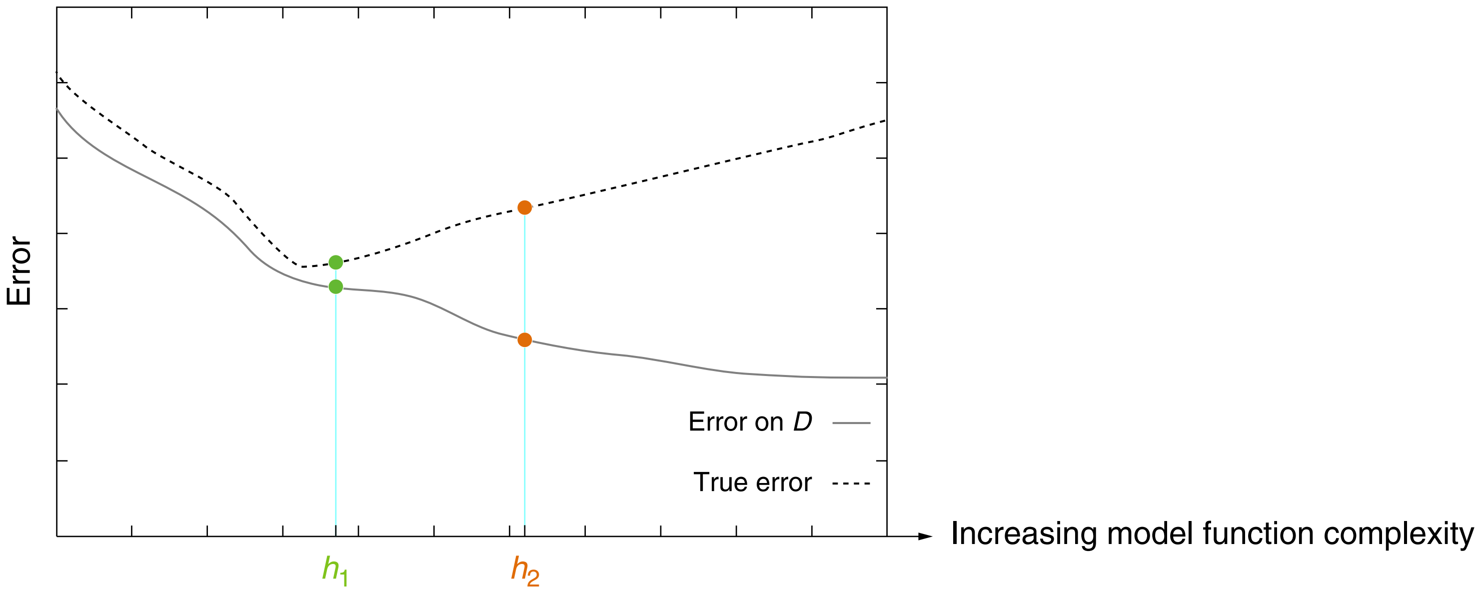
Definition 9 (Overfitting (continued)**)**

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$Err(h_2, D) < Err(h_1, D) \quad \text{and} \quad Err^*(h_1) < Err^*(h_2),$$

where $Err^*(h)$ denotes the true misclassification rate of h, while $Err(h, D)$ denotes the error of h for D.

[see first part]



Overfitting

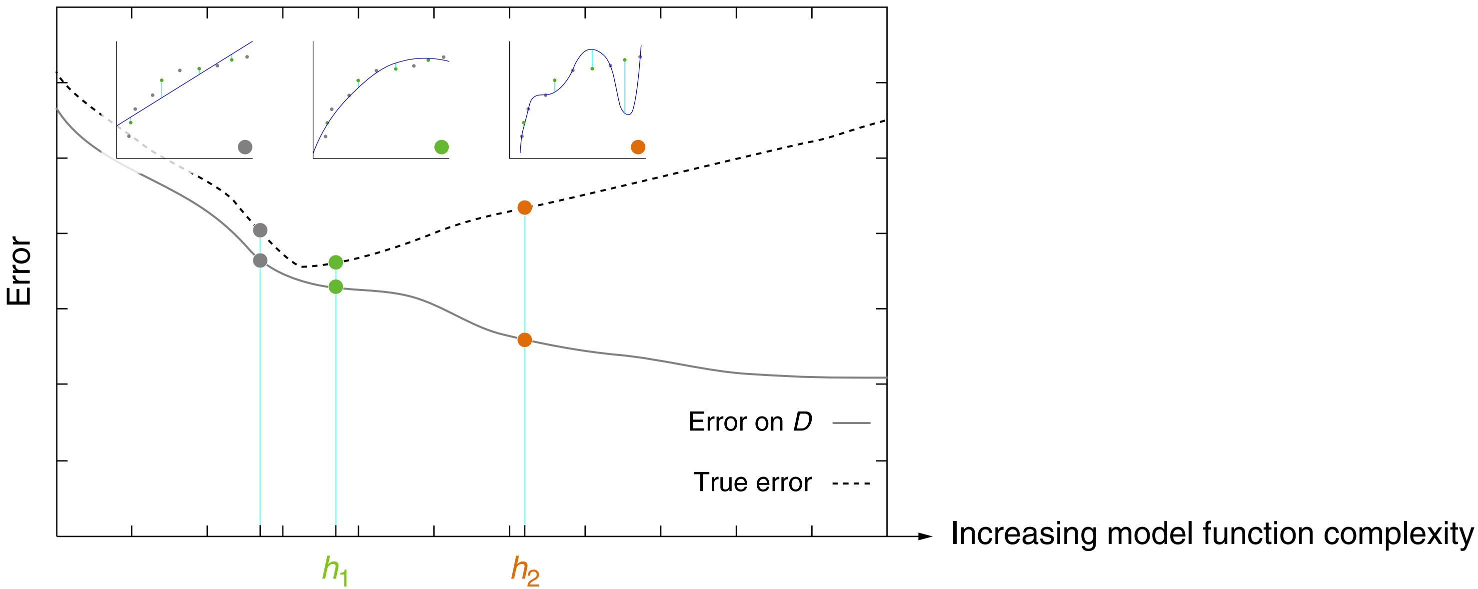
Definition 9 (Overfitting (continued)**)**

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$Err(h_2, D) < Err(h_1, D) \quad \text{and} \quad Err^*(h_1) < Err^*(h_2),$$

where $Err^*(h)$ denotes the true misclassification rate of h, while $Err(h, D)$ denotes the error of h for D.

[see first part]



Overfitting

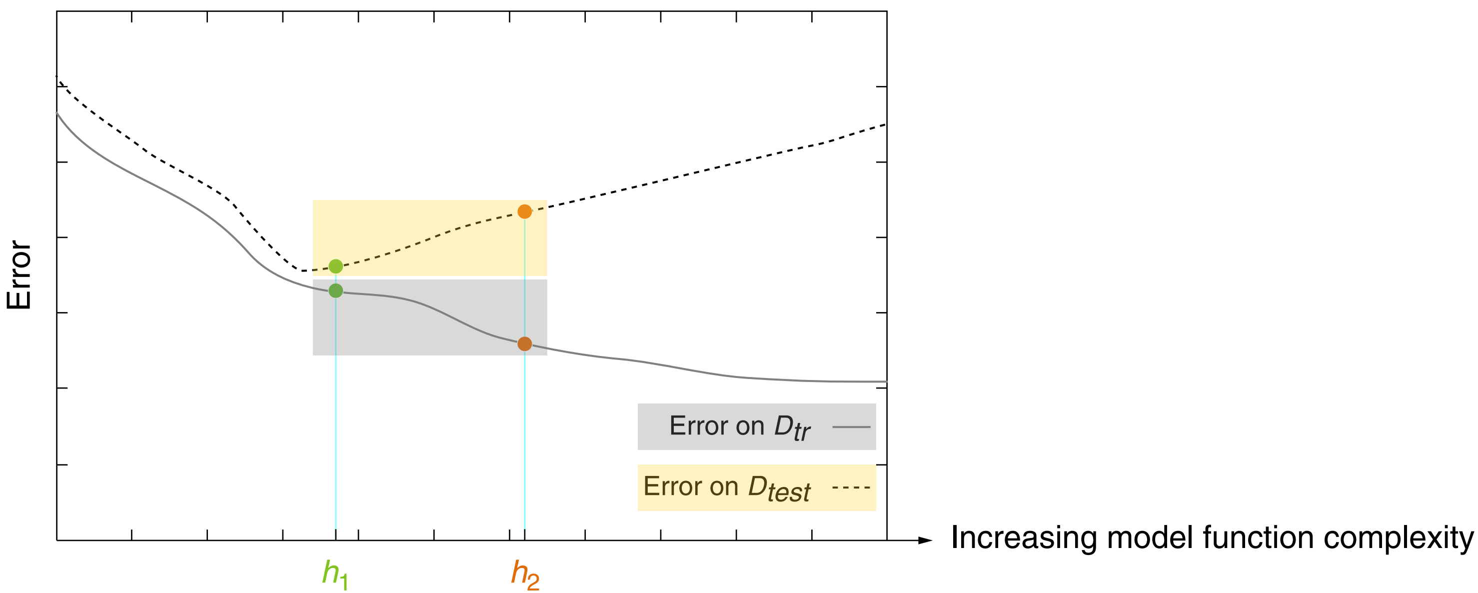
Definition 9 (Overfitting (continued)**)**

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h for D.

[see first part]



Overfitting

Definition 9 (Overfitting (continued))

Let D be a multiset of examples and let H be a hypothesis space. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D) < \mathit{Err}(h_1, D) \quad \text{and} \quad \mathit{Err}^*(h_1) < \mathit{Err}^*(h_2),$$

where $\mathit{Err}^*(h)$ denotes the true misclassification rate of h, while $\mathit{Err}(h, D)$ denotes the error of h for D.

Moreover, let $D_{tr} \subset D$ be a training set, $D_{test} \subset D$ be a test set, $D_{test} \cap D_{tr} = \emptyset$, and $\mathit{Err}(h, D_{test})$ be an estimate for $\mathit{Err}^*(h)$ [holdout estimation]. The hypothesis $h_2 \in H$ is considered to overfit D if an $h_1 \in H$ with the following property exists:

$$\mathit{Err}(h_2, D_{tr}) < \mathit{Err}(h_1, D_{tr}) \quad \text{and} \quad \mathit{Err}(h_1, D_{test}) < \mathit{Err}(h_2, D_{test})$$

In particular: $\mathit{Err}(h_2, D_{test}) \gg \mathit{Err}(h_2, D_{tr})$

Overfitting
Mitigation Strategies

How to detect overfitting:

- Visual inspection
  Apply projection or embedding for dimensionalities $p > 3$.

- Validation
  Given a test set, the difference $\mathit{Err}(y(), D_{\mathit{test}}) - \mathit{Err}(y(), D_{\mathit{tr}})$ is too large.

Overfitting

Mitigation Strategies (continued)

How to detect overfitting:

- Visual inspection
  Apply projection or embedding for dimensionalities $p > 3$.

- Validation
  Given a test set, the difference $\mathit{Err}(y(), D_{\mathit{test}}) - \mathit{Err}(y(), D_{\mathit{tr}})$ is too large.

How to tackle overfitting:

- Increase the quantity and / or the quality of the training data D.
  Quantity: More data averages out noise.
  Quality: Omitting "poor examples" allows a better fit, but is problematic though.

- Early stopping of the optimization (training) process.
  Criterion: $\mathit{Err}(y(), D_{\mathit{test}}) - \mathit{Err}(y(), D_{\mathit{tr}})$ increases with the number of iterations (training time).

Overfitting

Mitigation Strategies (continued)

How to detect overfitting:

- Visual inspection
  Apply projection or embedding for dimensionalities $p > 3$.

- Validation
  Given a test set, the difference $\mathit{Err}(y(), D_{\mathit{test}}) - \mathit{Err}(y(), D_{\mathit{tr}})$ is too large.

How to tackle overfitting:

- Increase the quantity and / or the quality of the training data D.
  Quantity: More data averages out noise.
  Quality: Omitting “poor examples” allows a better fit, but is problematic though.

- Early stopping of the optimization (training) process.
  Criterion: $\mathit{Err}(y(), D_{\mathit{test}}) - \mathit{Err}(y(), D_{\mathit{tr}})$ increases with the number of iterations (training time).

- Regularization: Increase model bias by constraining the hypothesis space.
  (1) Model function: Consider functions of lower complexity / VC dimension. [Wikipedia]
  (2) Hypothesis $\mathbf{w}$: Bound the absolute values of the weights in $\vec{\mathbf{w}}$ of a model function.

Chapter ML:III (continued)

III. Linear Models

- Logistic Regression
- Loss Computation in Detail
- Overfitting
- Regularization
- Gradient Descent in Detail

Regularization

Regularization in the Machine Learning Stack

Optimization approach

Optimization objective

Loss function [+ Regularization]

Model function $\rightsquigarrow$ Hypothesis space

Task

Data

⋮

- Objective: minimize objective function
- Regularization: Euclidean norm, absolute-value norm
- Loss: $l(c, y(\mathbf{x}))$

⋮

⋮

Regularization

Bound the Absolute Values of the Weights $\mathbf{w}$

Principle: Add to the loss function (term) a regularization function (term), $R(\mathbf{w})$:

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R(\mathbf{w}), \qquad \ell(\mathbf{w}) \;=\; l(c, y(\mathbf{x})) \;+\; \frac{\lambda}{n} \cdot R(\mathbf{w}),$$

where $\lambda \geq 0$ controls the impact of $R(\mathbf{w})$, $R(\mathbf{w}) \geq 0$.

Regularization

Bound the Absolute Values of the Weights $\mathbf{w}$ (continued)

Principle: Add to the loss function (term) a regularization function (term), $R(\mathbf{w})$:

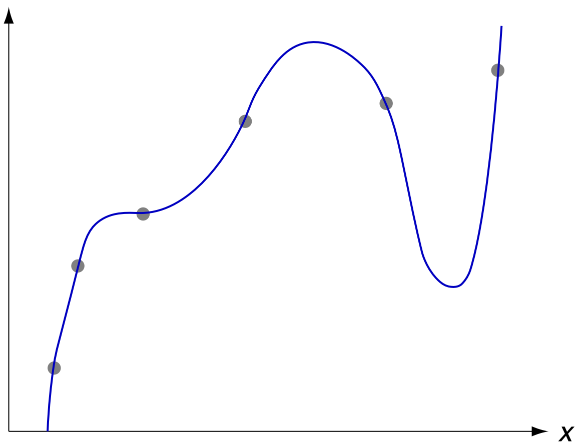
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R(\mathbf{w}), \qquad \ell(\mathbf{w}) \;=\; l(c, y(\mathbf{x})) \;+\; \frac{\lambda}{n} \cdot R(\mathbf{w}),$$

where $\lambda \geq 0$ controls the impact of $R(\mathbf{w})$, $R(\mathbf{w}) \geq 0$.

Example (c) (continued) :

- $y(x) = w_0 + \sum_{j=1}^{6} w_j \cdot x^j$
- $L(\mathbf{w}) = \text{RSS}(\mathbf{w}) = \sum_{i=1}^{n} (y_i - y(x_i))^2$
- $R(\mathbf{w}) = |w_1| + |w_2| + \ldots + |w_6|$
  $\lambda = 0$

$\leadsto$ $\hat{\mathbf{w}} = (-0.7,\ 15.4,\ -80.6,\ 174.9,\ -99.5,\ -113.7,\ 109.7)^T$



Regularization

Bound the Absolute Values of the Weights $\mathbf{w}$ (continued)

Principle: Add to the loss function (term) a regularization function (term), $R(\mathbf{w})$:

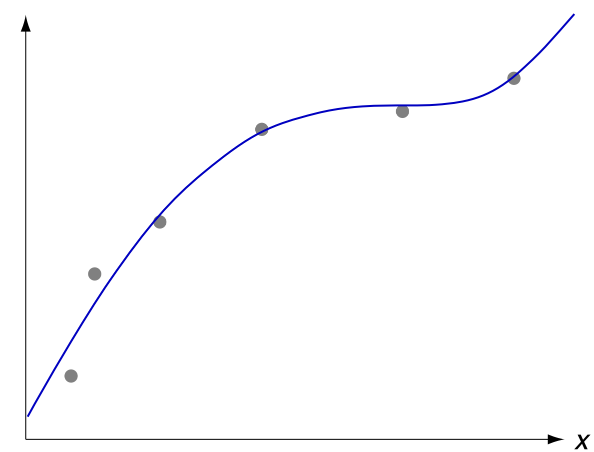
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R(\mathbf{w}), \qquad \ell(\mathbf{w}) \;=\; l(c, y(\mathbf{x})) \;+\; \frac{\lambda}{n} \cdot R(\mathbf{w}),$$

where $\lambda \geq 0$ controls the impact of $R(\mathbf{w})$, $R(\mathbf{w}) \geq 0$.

Example (c) (continued) :

- $y(x) = w_0 + \sum_{j=1}^{6} w_j \cdot x^j$
- $L(\mathbf{w}) = \text{RSS}(\mathbf{w}) = \sum_{i=1}^{n} (y_i - y(x_i))^2$
- $R(\mathbf{w}) = |w_1| + |w_2| + \ldots + |w_6|$
  $\lambda = 0.001$

$\rightsquigarrow$ $\hat{\mathbf{w}} = (0.01,\ 2.0,\ -1.73,\ -0.22,\ 0.0,\ 0.0,\ 0.8)^T$



Regularization

Bound the Absolute Values of the Weights $\mathbf{w}$ (continued)

Principle: Add to the loss function (term) a regularization function (term), $R(\mathbf{w})$:

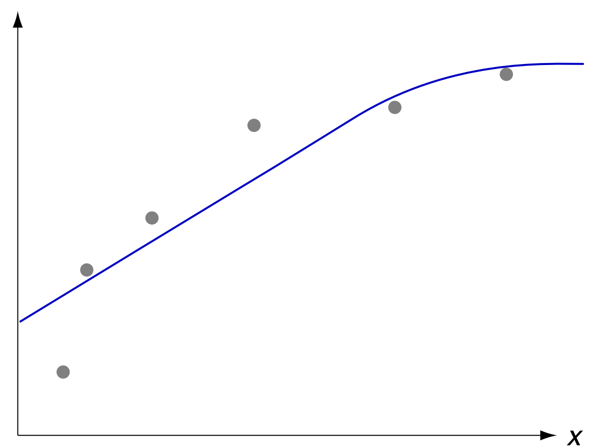
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R(\mathbf{w}), \qquad \ell(\mathbf{w}) \;=\; l(c, y(\mathbf{x})) \;+\; \frac{\lambda}{n} \cdot R(\mathbf{w}),$$

where $\lambda \geq 0$ controls the impact of $R(\mathbf{w})$, $R(\mathbf{w}) \geq 0$.

Example (c) (continued) :

- $y(x) = w_0 + \sum_{j=1}^{6} w_j \cdot x^j$
- $L(\mathbf{w}) = \text{RSS}(\mathbf{w}) = \sum_{i=1}^{n} (y_i - y(x_i))^2$
- $R(\mathbf{w}) = |w_1| + |w_2| + \ldots + |w_6|$
  $\lambda = 0.02$

$\rightsquigarrow$ $\hat{\mathbf{w}} = (0.17,\ 0.73,\ 0.0,\ -0.21,\ -0.01,\ -0.01,\ 0.0)^T$



Regularization

Bound the Absolute Values of the Weights $\mathbf{w}$ (continued)

Principle: Add to the loss function (term) a regularization function (term), $R(\mathbf{w})$:

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R(\mathbf{w}), \qquad \ell(\mathbf{w}) \;=\; l(c, y(\mathbf{x})) \;+\; \frac{\lambda}{n} \cdot R(\mathbf{w}),$$

where $\lambda \geq 0$ controls the impact of $R(\mathbf{w})$, $R(\mathbf{w}) \geq 0$.

Observations:

- Model complexity depends (also) on the magnitude of the weights $\mathbf{w}$.
- Minimizing $L(\mathbf{w})$ sets no bounds on the weights $\mathbf{w}$.
- Regularization is achieved with a “counterweight” $\lambda \cdot R(\mathbf{w})$ that grows with $\mathbf{w}$.
- Aside from λ no additional hyperparameter is introduced.

Remarks:

- $\mathcal{L}(\mathbf{w})$ is called (global) "objective function", "cost function", or "error function"; $\ell(\mathbf{w})$ is its pointwise counterpart.

- The regularization term constrains the magnitude of the direction vector of the hyperplane, progressively reducing the hyperplane's steepness as λ increases. The intercept w_0 is adjusted accordingly through minimization of $\mathcal{L}(\mathbf{w})$ but must not be part of the regularization term itself, which would lead to an incorrect solution.

- To denote the difference, we write $\mathbf{w} \equiv (w_0, w_1, \ldots, w_p)^T$ to refer to the entire parameter vector (the actual hypothesis), and $\vec{\mathbf{w}} \equiv (w_1, \ldots, w_p)^T$ for the direction vector excluding w_0.

- About choosing λ:
  - Recall subsection Comparing Model Variants of section Evaluating Effectiveness where hyperparameter optimization is tackled by means of a validation set.
  - How to calculate the regularization parameter λ in linear regression. [stackoverflow]
  - "No black-box procedures for choosing the regularization parameter λ are available, and most likely will never exist." [Hansen/Hanke 1993]

Remarks (continued) :

- The term “regularization” derives from “regular”, a synonym for “smooth” within the context of model functions. [stackexchange]

- Regularization is applied in settings where the set of examples D is much smaller than the population of real-world objects O. Under the conditions of the Inductive Learning Hypothesis we can infer from D a hypothesis h that generalizes sufficiently well to the entire population—if h is sufficiently simple, stable (wrt. changes in D), and smooth, which can be reached with regularization.

  However, if D covers (nearly) the entire population, minimizing the loss $L(\mathbf{w})$ takes precedence over additional restrictions $R(\mathbf{w})$ regarding the simplicity, the stability, and the smoothness of h.

- The origins of regularization go back to the fields of inverse problems and ill-posed problems. Solving an inverse problem means calculating from a set of observations the causal factors that produced them. [Wikipedia]

  Inverse problems are often ill-posed, where “ill-posedness” is defined as not being “well-posed”. In turn, a mathematical problem is called well-posed if (1) a solution exists, (2) the solution is unique, (3) the solution’s behavior changes continuously with the initial conditions. [Wikipedia]

  Under certain assumptions the problem of learning from examples forms an inverse problem. [deVito 2005]

Regularization

The Vector Norm as Regularization Function

- Ridge regression. $$R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w}) \;=\; \sum_{i=1}^{p} w_i^2 \;=\; \vec{\mathbf{w}}^T\vec{\mathbf{w}}$$

- Lasso regression. $$R_{||\vec{\mathbf{w}}||_1}(\mathbf{w}) \;=\; \sum_{i=1}^{p} |w_i|$$

Remarks:

- The term “ridge” refers to the ridge that one gets in the likelihood function (equivalently, “valley” in the RSS) if the there is multicollinearity in the data. Ridge regression adds a penalty that turns the ridge into a peak in likelihood space or, equivalently, a depression in the minimization criterion. [stackexchange]

  Ridge regression predates lasso regression. It is also known as weight decay in machine learning, and with multiple independent discoveries, it is variously known as the Tikhonov-Miller method, the Phillips-Twomey method, the constrained linear inversion method, and the method of linear regularization. [Wikipedia]

- “Lasso” is an acronym for “least absolute shrinkage and selection operator”.

- $||\cdot||_k$ denotes the vector norm operator:

$$||\mathbf{x}||_k \equiv \left(\sum_{j=1}^{p} |x_j|^k\right)^{1/k},$$

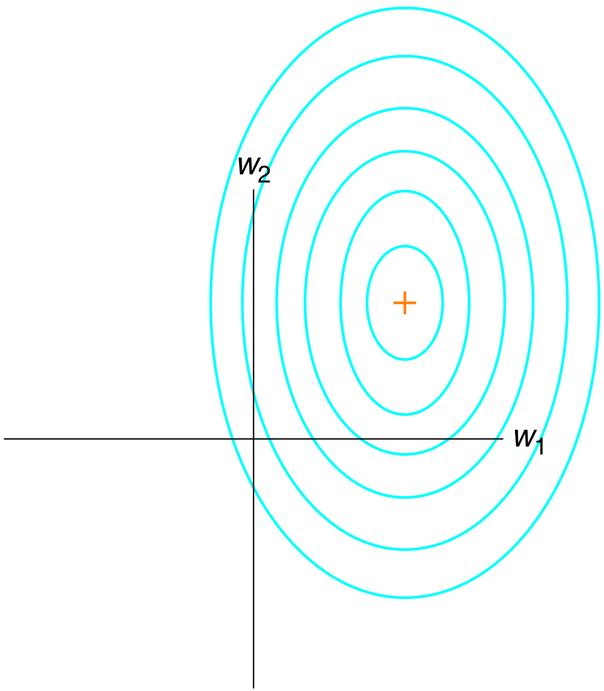
  where $k \in [1, \infty)$ and p is the dimensionality of vector $\mathbf{x}$.

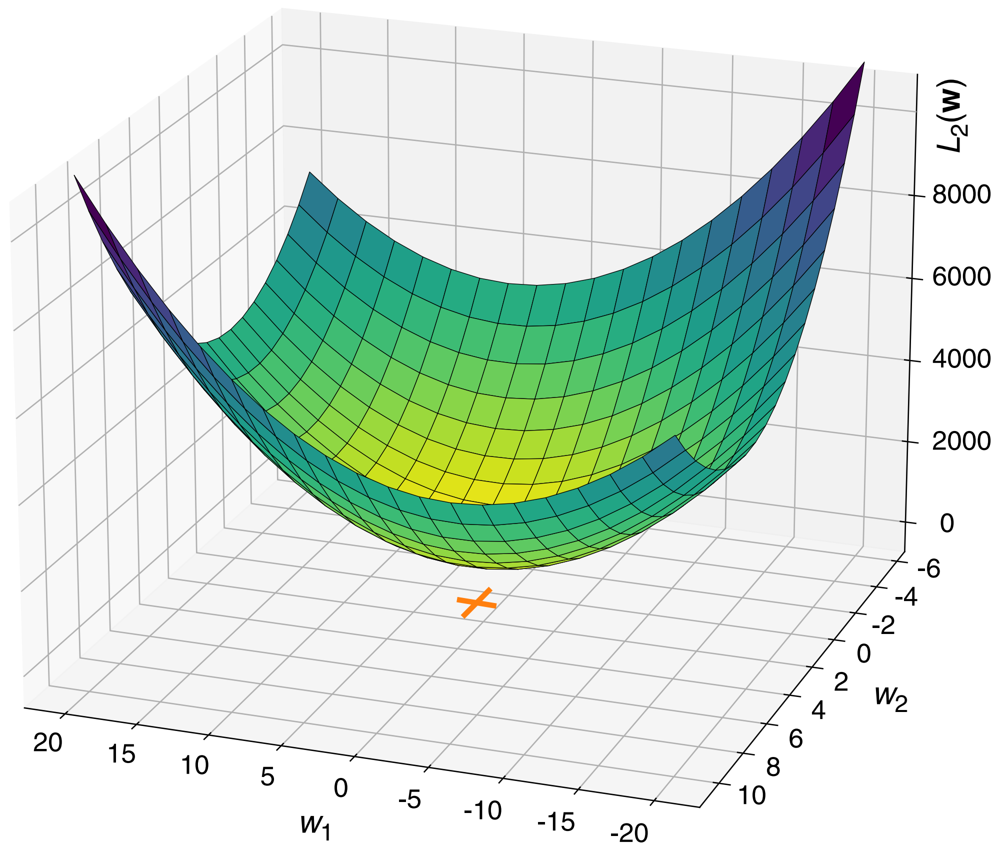
- By convention, $||\cdot||$ (omitting the subscript) refers to the Euclidean norm ($k = 2$).

Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) = L(\mathbf{w})$$

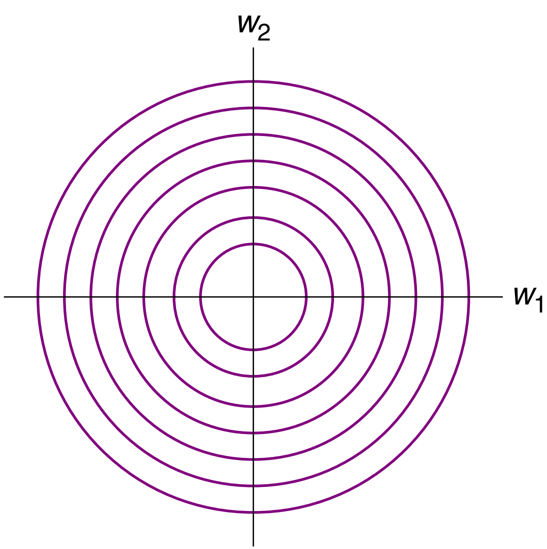


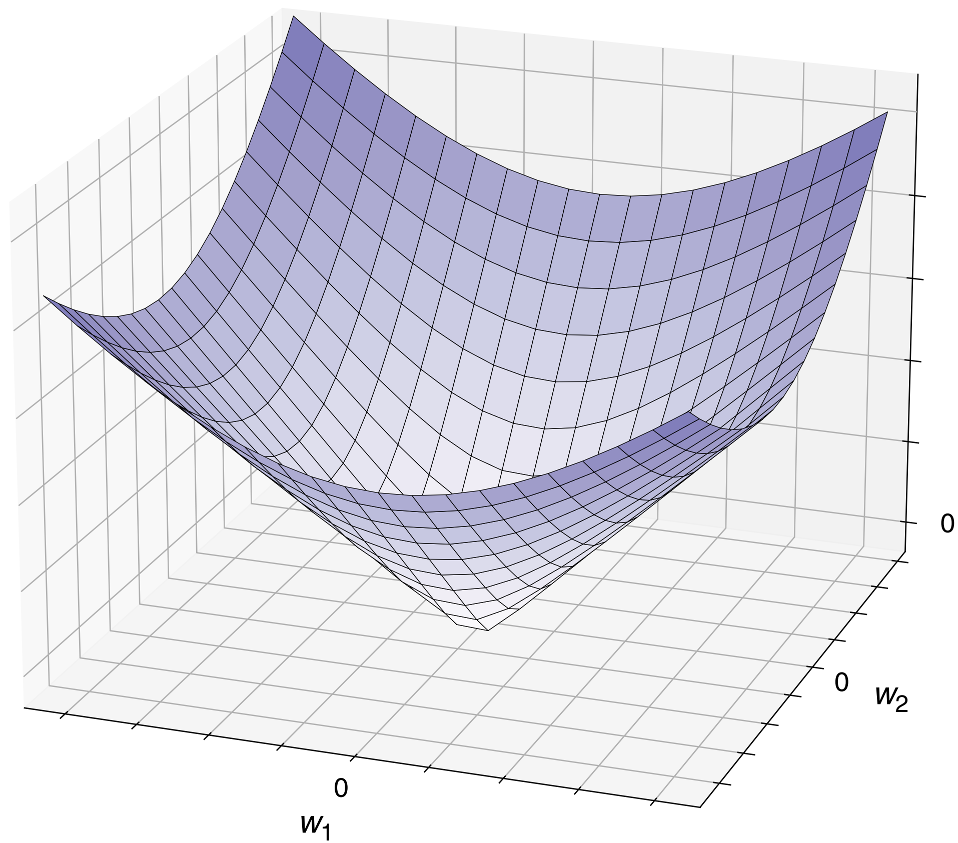


Regularization

The Vector Norm as Regularization Function (continued)

$$R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w}) \;=\; \sum_{i=1}^{p} w_i^2 \;=\; \vec{\mathbf{w}}^T\vec{\mathbf{w}}$$

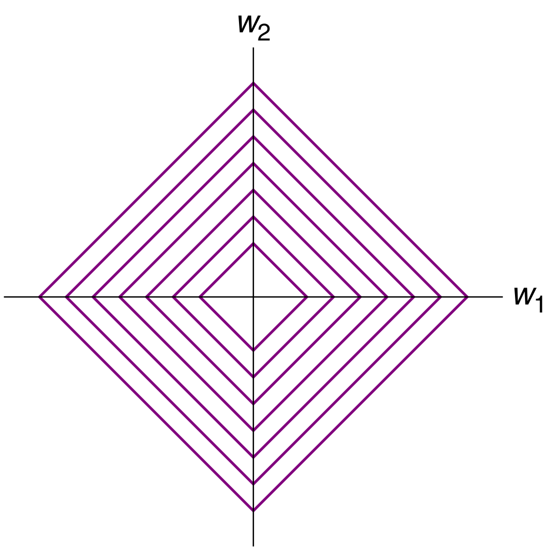


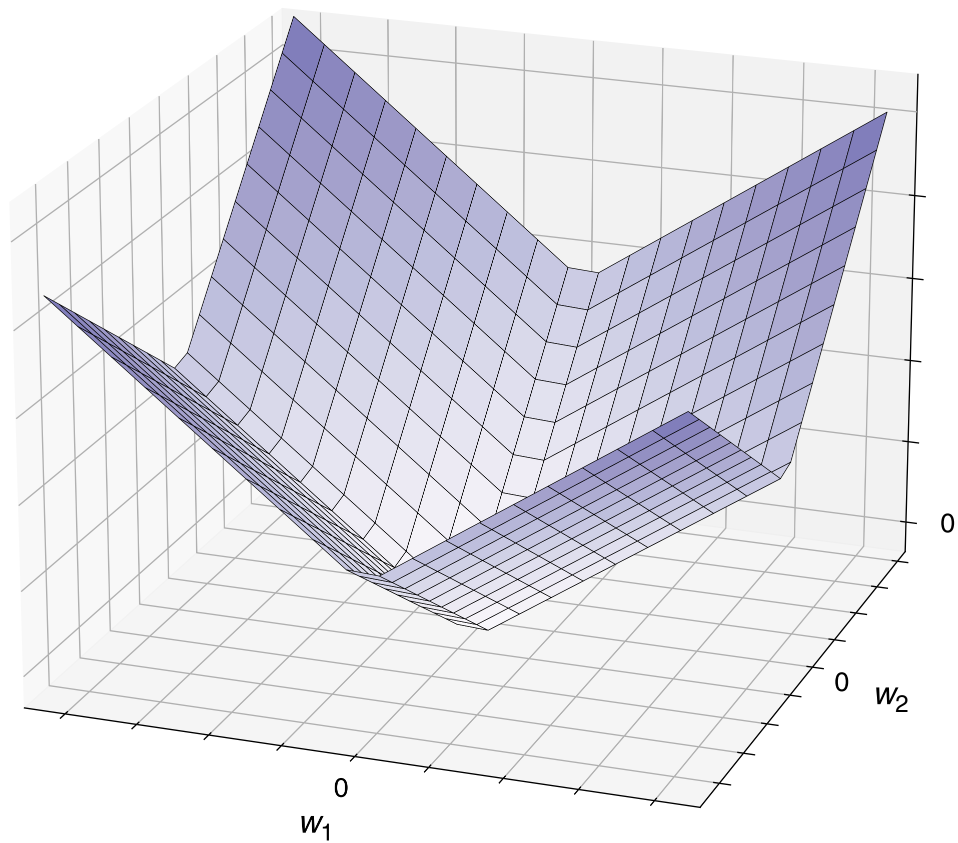


Regularization

The Vector Norm as Regularization Function (continued)

$$R_{||\vec{\mathbf{w}}||_1}(\mathbf{w}) \;=\; \sum_{i=1}^{p} |w_i|$$





Remarks:

- The exemplified plots of the loss term, $L(\mathbf{w})$, and the regularization term, $R(\mathbf{w})$, are illustrated over the parameter space $\{(w_1, w_2) \mid w_i \in \mathbf{R}\}$ (instead of $\{(w_0, w_1) \mid w_i \in \mathbf{R}\}$) to better emphasize the characteristic difference between ridge regression and lasso regression.

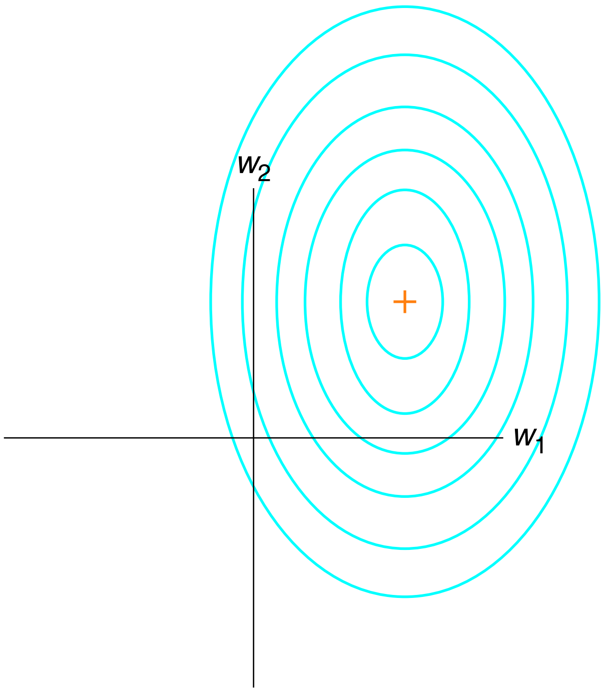
- The contour line plots show two-dimensional projections of the three-dimensional convex loss function (here: RSS) for a given set of example data, as well as of the two regularization functions $R_{||\mathbf{w}||_2^2}$ and $R_{||\mathbf{w}||_1}$, whose shapes do not depend on the data.

- A contour line is a curve along which the respective function has a constant value.

Regularization

The Vector Norm as Regularization Function (continued)

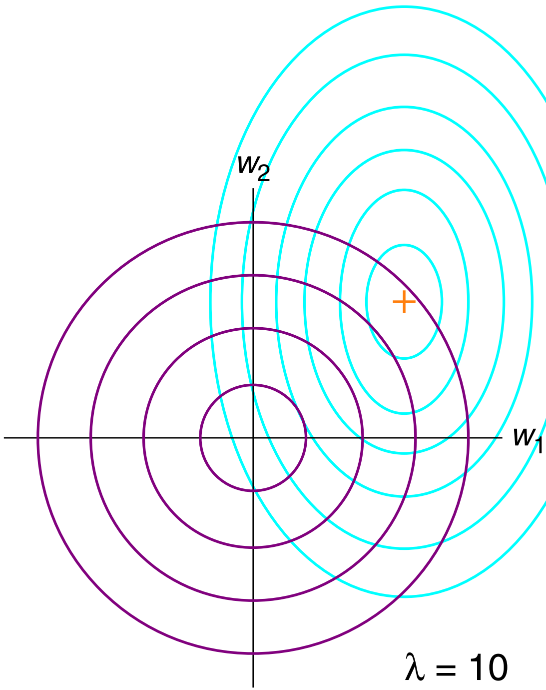
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w})$$



Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$

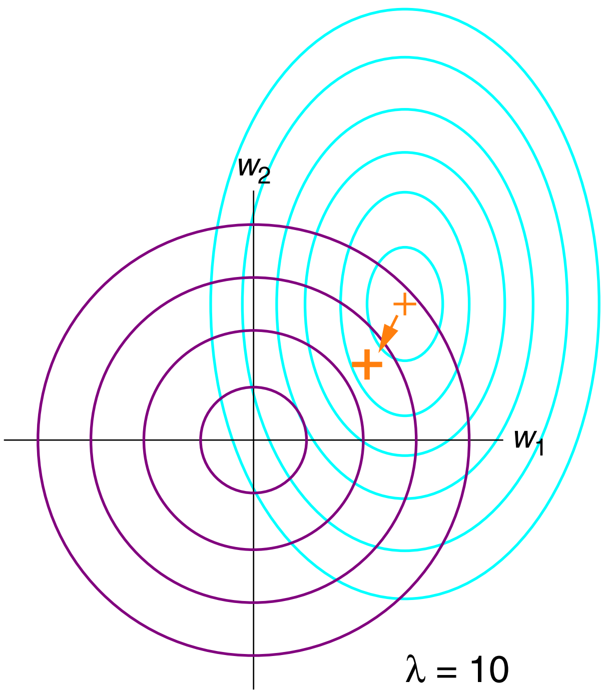


$\lambda = 10$

Regularization

The Vector Norm as Regularization Function (continued)

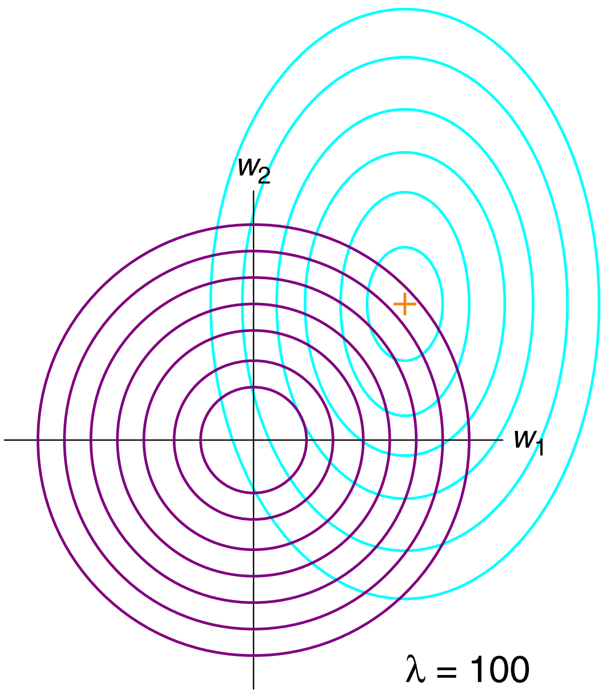
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$



Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \; = \; L(\mathbf{w}) \; + \; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$

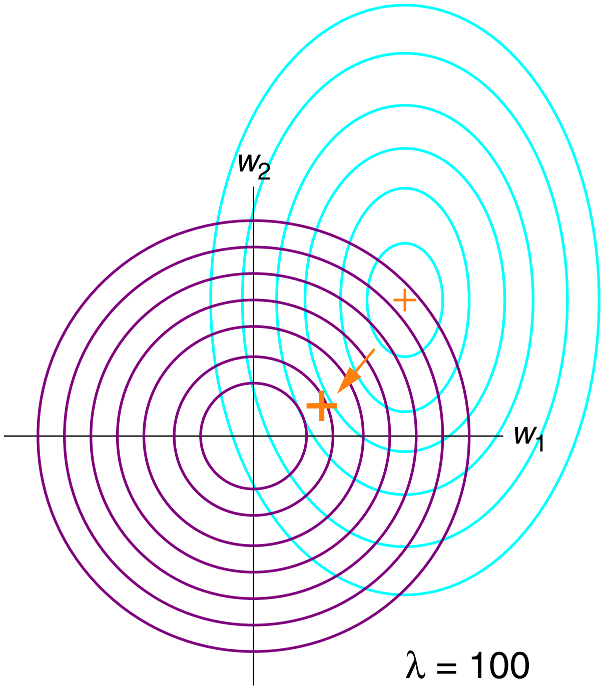


$\lambda = 100$

Regularization

The Vector Norm as Regularization Function (continued)

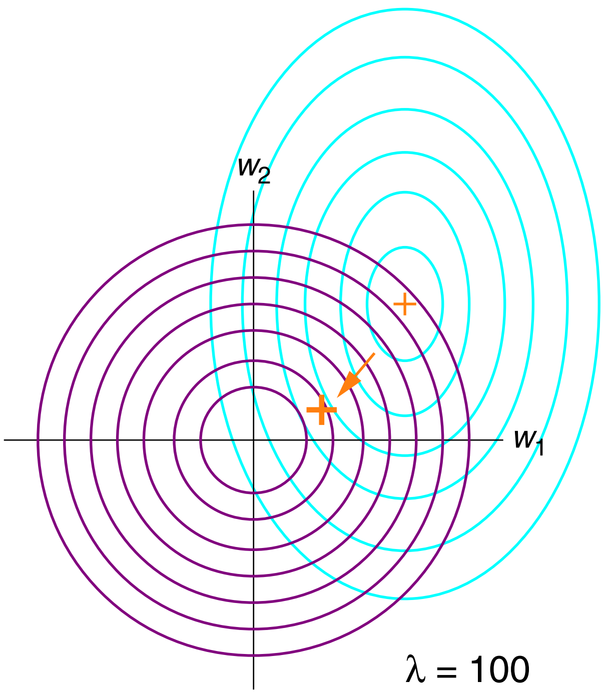
$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$



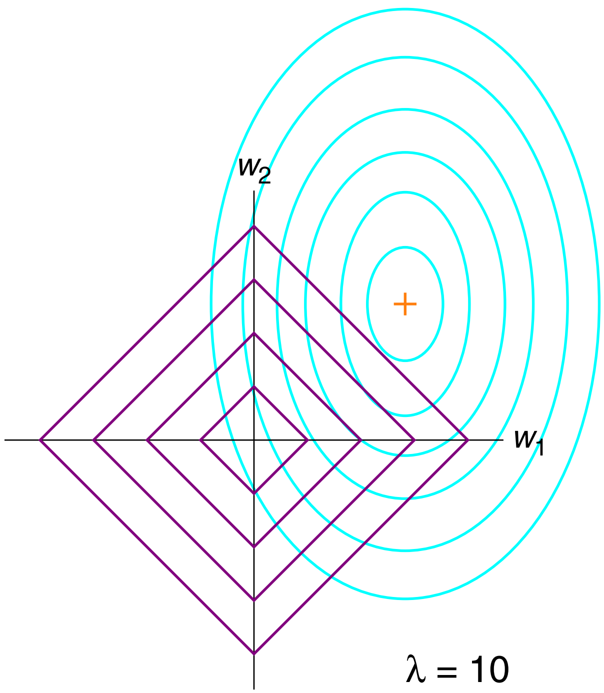
Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$



$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_1}(\mathbf{w})$$

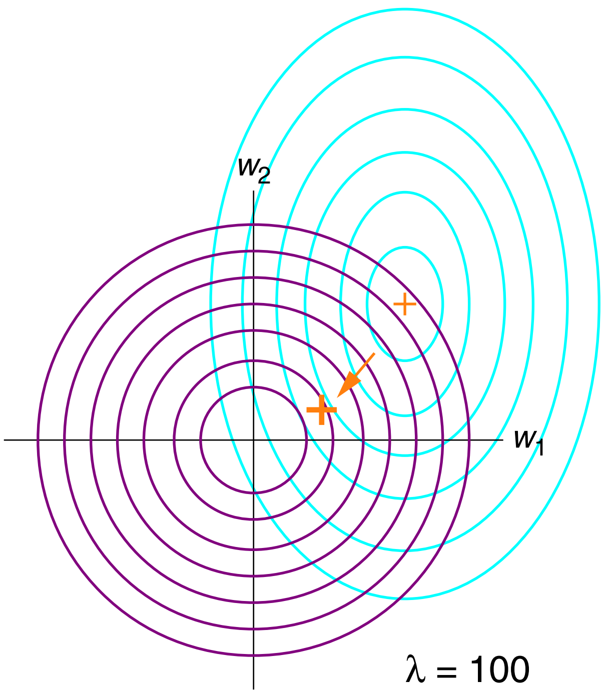


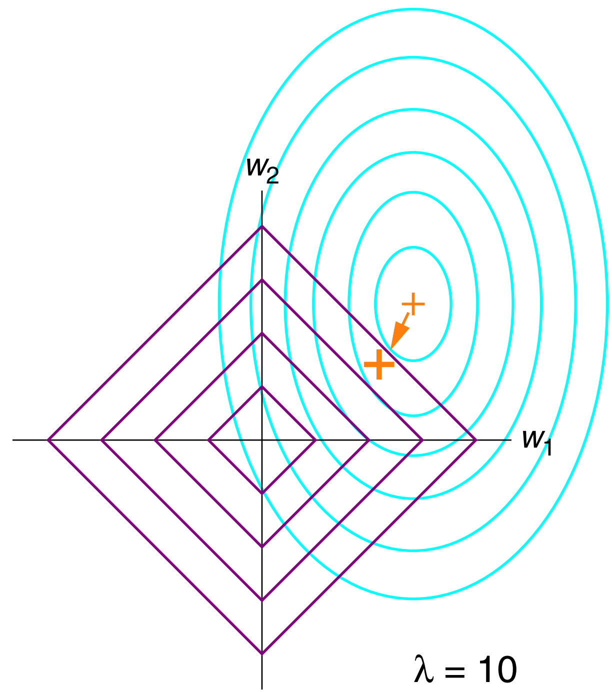
Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_1}(\mathbf{w})$$



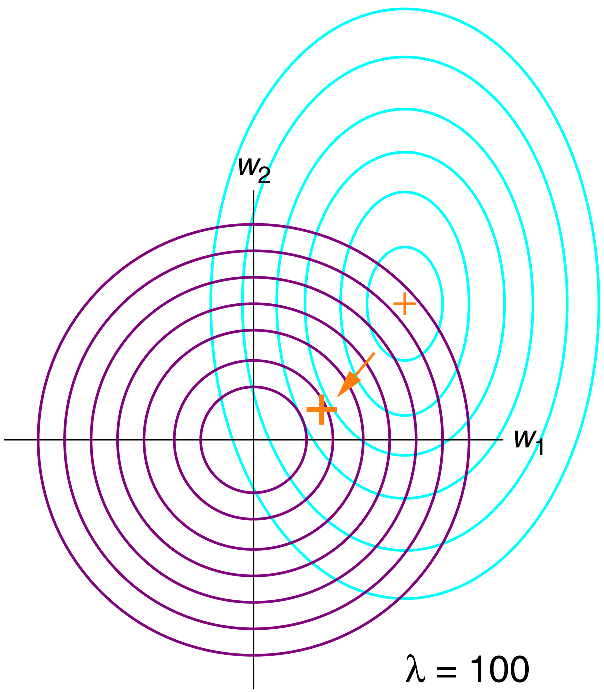


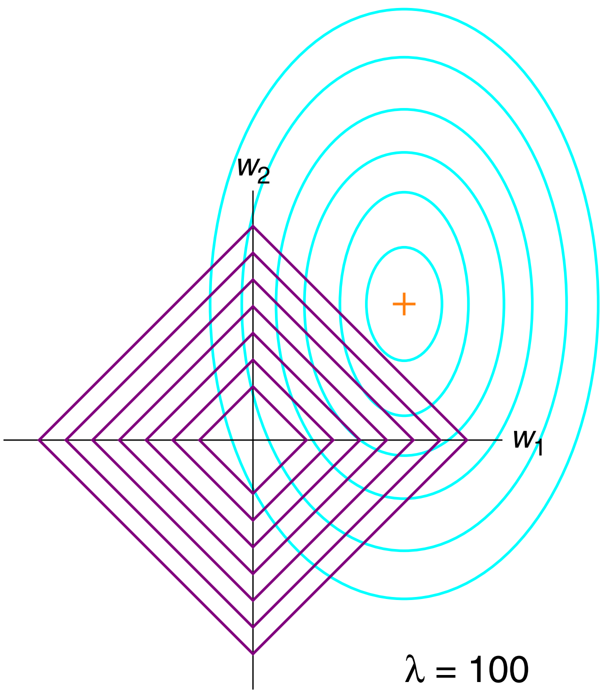
Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$

$$\mathcal{L}(\mathbf{w}) \;=\; L(\mathbf{w}) \;+\; \lambda \cdot R_{||\vec{\mathbf{w}}||_1}(\mathbf{w})$$

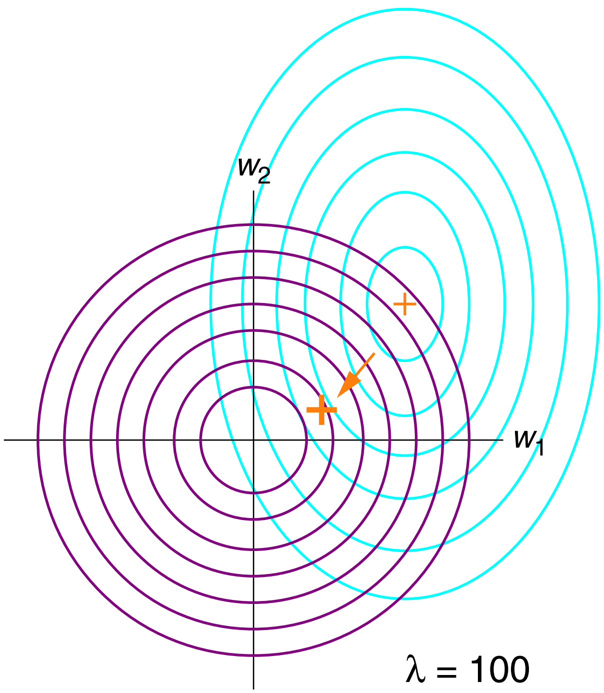




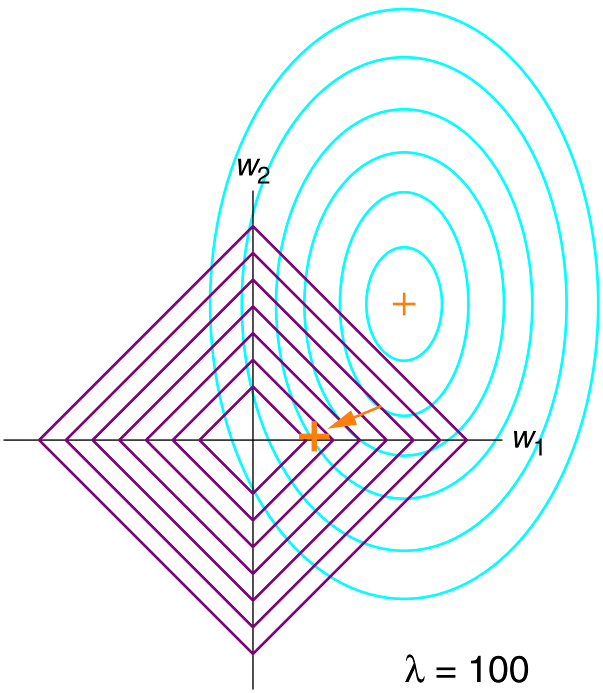
Regularization

The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) \ = \ L(\mathbf{w}) \ + \ \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$



$$\mathcal{L}(\mathbf{w}) \ = \ L(\mathbf{w}) \ + \ \lambda \cdot R_{||\vec{\mathbf{w}}||_1}(\mathbf{w})$$

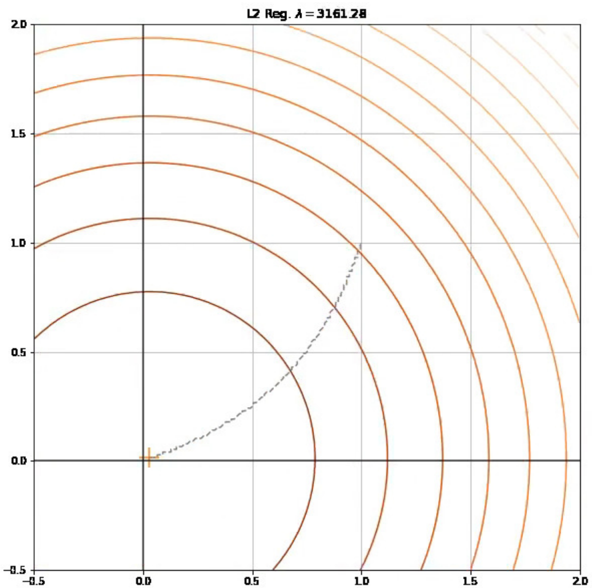


Regularization

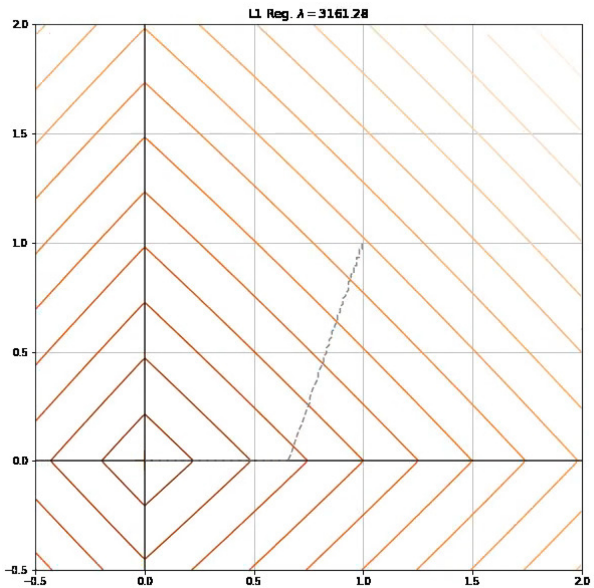
The Vector Norm as Regularization Function (continued)

$$\mathcal{L}(\mathbf{w}) = L(\mathbf{w}) + \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w})$$

$$\mathcal{L}(\mathbf{w}) = L(\mathbf{w}) + \lambda \cdot R_{||\vec{\mathbf{w}}||_1}(\mathbf{w})$$



[animation]



[animation]

The animations show superimposed contourlines. The choice of R determines the trajectory the minimum takes towards the origin as a function of λ. [stackexchange]

Remarks:

- The exemplified loss function is minimal at the cross. Without regularization, the weights associated with the minimum will be the result of a linear regression. By adding the regularization term $\lambda \cdot R(\mathbf{w})$ with $\lambda > 0$, the joint minimum of the two functions is found closer to the origin of the parameter space than the minimum of the loss function.

- The choice of λ determines how much closer the joint minimum is to the origin of the parameter space; the higher, the closer, and thus the smaller the parameters $\mathbf{w}$.

- The minimum of $\mathcal{L}(\mathbf{w})$ is on a tangent point between a contour line of $L(\mathbf{w})$ and a contour line of $R(\mathbf{w})$. Barring exceptional cases, the minimum of $\mathcal{L}(\mathbf{w})$ (the sum of global loss and regularization) is unique, even if the minimum of $L(\mathbf{w})$ (the global loss) is non-unique.

- A key difference between ridge ($R_{||\vec{\mathbf{w}}||_2^2}$) and lasso ($R_{||\vec{\mathbf{w}}||_1}$) regression is that, with lasso regression, parameters can be reduced to zero, eliminating the corresponding feature from the model function.

  With ridge regression, the influence of all parameters will be reduced "uniformly." In particular, a parameter will be reduced to zero if and only if the minimum of the loss function is found on that parameter's axis.

Regularization

Regularized Linear Regression [linear regression]

- Given $\mathbf{x}$, predict a real-valued output under a linear model function:

$$y(\mathbf{x}) = w_0 + \sum_{j=1}^{p} w_j \cdot x_j$$

- Vector notation with $x_0 = 1$ and $\mathbf{w} = (w_0, w_1, \ldots, w_p)^T$:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

Regularization

Regularized Linear Regression (continued) [linear regression]

- Given $\mathbf{x}$, predict a real-valued output under a linear model function:

$$y(\mathbf{x}) = w_0 + \sum_{j=1}^{p} w_j \cdot x_j$$

- Vector notation with $x_0 = 1$ and $\mathbf{w} = (w_0, w_1, \ldots, w_p)^T$:

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

- Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, assess goodness of fit of the objective function:

$$\mathcal{L}(\mathbf{w}) \; = \; \text{RSS}(\mathbf{w}) + \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w}) \; = \; \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \cdot \vec{\mathbf{w}}^T \vec{\mathbf{w}} \qquad (1)$$

Regularization

Regularized Linear Regression (continued) [linear regression]

- Given $\mathbf{x}$, predict a real-valued output under a linear model function:

$$y(\mathbf{x}) = w_0 + \sum_{j=1}^{p} w_j \cdot x_j$$

- Vector notation with $x_0 = 1$ and $\mathbf{w} = (w_0, w_1, \ldots, w_p)^T$:

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

- Given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, assess goodness of fit of the objective function:

$$\mathcal{L}(\mathbf{w}) \;=\; \text{RSS}(\mathbf{w}) + \lambda \cdot R_{||\vec{\mathbf{w}}||_2^2}(\mathbf{w}) \;=\; \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \cdot \vec{\mathbf{w}}^T \vec{\mathbf{w}} \qquad (1)$$

- Estimate optimum $\mathbf{w}$ by minimizing the residual sum of squares:

$$\hat{\mathbf{w}} \;=\; \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\text{argmin}} \; \mathcal{L}(\mathbf{w}) \qquad (2)$$

Regularization

Regularized Linear Regression (continued) [linear regression]

- Let X denote the $n \times (p+1)$ matrix, where row i is $(1 \;\; \mathbf{x}_i^T)$ with $(\mathbf{x}_i, y_i) \in D$.

  Let $\mathbf{y}$ denote the n-vector of outputs in the training set D.

$\rightsquigarrow \quad \mathcal{L}(\mathbf{w}) \;=\; (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) \;+\; \lambda \cdot \vec{\mathbf{w}}^T\vec{\mathbf{w}}$

Regularized Linear Regression (continued) [linear regression]

- Let X denote the $n \times (p{+}1)$ matrix, where row i is $(1 \;\; \mathbf{x}_i^T)$ with $(\mathbf{x}_i, y_i) \in D$.

  Let $\mathbf{y}$ denote the n-vector of outputs in the training set D.

$$\leadsto \quad \mathcal{L}(\mathbf{w}) = (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \lambda \cdot \vec{\mathbf{w}}^T\vec{\mathbf{w}}$$

- Minimize $\mathcal{L}(\mathbf{w})$ via a direct method:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} &= -2X^T(\mathbf{y} - X\mathbf{w}) + 2\lambda \cdot \binom{0}{\vec{\mathbf{w}}} = 0 \\
X^T(\mathbf{y} - X\mathbf{w}) - \lambda \cdot \binom{0}{\vec{\mathbf{w}}} &= 0 \\
\Leftrightarrow \quad \left(X^T X + \lambda \cdot \text{diag}(0, 1, \ldots, 1)\right) \mathbf{w} &= X^T\mathbf{y} \\
\Leftrightarrow \quad \mathbf{w} &= \left(X^T X + \text{diag}(0, \lambda, \ldots, \lambda) \right)^{-1} X^T\mathbf{y}
\end{aligned}$$

Regularization

Regularized Linear Regression (continued) [linear regression]

- Let X denote the $n \times (p{+}1)$ matrix, where row i is $(1 \;\; \mathbf{x}_i^T)$ with $(\mathbf{x}_i, y_i) \in D$.

  Let $\mathbf{y}$ denote the n-vector of outputs in the training set D.

$$\rightsquigarrow \quad \mathcal{L}(\mathbf{w}) \;=\; (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) \;+\; \lambda \cdot \vec{\mathbf{w}}^T\vec{\mathbf{w}}$$

- Minimize $\mathcal{L}(\mathbf{w})$ via a direct method:

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} \;=\; -2X^T(\mathbf{y} - X\mathbf{w}) \;+\; 2\lambda \cdot \binom{0}{\vec{\mathbf{w}}} \;=\; 0$$

$$X^T(\mathbf{y} - X\mathbf{w}) - \lambda \cdot \binom{0}{\vec{\mathbf{w}}} \;=\; 0$$

$$\Leftrightarrow \quad \left(X^T X + \lambda \cdot \text{diag}(0, 1, \ldots, 1)\right) \mathbf{w} \;=\; X^T\mathbf{y} \qquad \text{Normal equations.}$$

$$\Leftrightarrow \quad \mathbf{w} \;=\; \big(\underbrace{X^T X + \text{diag}(0, \lambda, \ldots, \lambda)}_{\text{Conditioning the moment matrix } X^T X}\big)^{-1} X^T\mathbf{y} \quad \text{If } \lambda > 0.$$

Conditioning the moment matrix $X^T X$ [Wikipedia 1, 2, 3]

Regularization

Regularized Linear Regression (continued) [linear regression]

- Let X denote the $n \times (p+1)$ matrix, where row i is $(1 \;\; \mathbf{x}_i^T)$ with $(\mathbf{x}_i, y_i) \in D$.

  Let $\mathbf{y}$ denote the n-vector of outputs in the training set D.

$$\rightsquigarrow \quad \mathcal{L}(\mathbf{w}) \;=\; (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) \;+\; \lambda \cdot \vec{\mathbf{w}}^T\vec{\mathbf{w}}$$

- Minimize $\mathcal{L}(\mathbf{w})$ via a direct method:

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} \;=\; -2X^T(\mathbf{y} - X\mathbf{w}) \;+\; 2\lambda \cdot \binom{0}{\vec{\mathbf{w}}} \;=\; 0$$

$$X^T(\mathbf{y} - X\mathbf{w}) - \lambda \cdot \binom{0}{\vec{\mathbf{w}}} \;=\; 0$$

$$\Leftrightarrow \quad \left(X^T X + \lambda \cdot \text{diag}(0, 1, \ldots, 1)\right) \mathbf{w} \;=\; X^T\mathbf{y}$$

Normal equations.

$$\Leftrightarrow \quad \hat{\mathbf{w}} \equiv \mathbf{w} \;=\; \Big(\underbrace{X^T X + \text{diag}(0, \lambda, \ldots, \lambda)}_{}\Big)^{-1} X^T\mathbf{y} \text{If } \lambda > 0.$$

Conditioning the moment matrix X^TX [Wikipedia 1, 2, 3]

$$\hat{y}(\mathbf{x}_i) \;=\; \hat{\mathbf{w}}^T \, \mathbf{x}_i$$

Regression function with least squares estimator $\hat{\mathbf{w}}$.