Lecture 3: Linear Regression
CS480/680 Intro to Machine Learning

2023-1-17

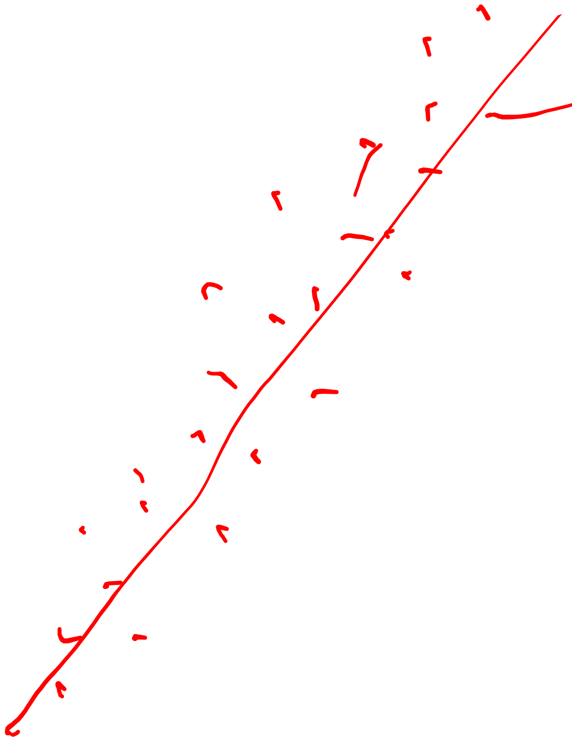
Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Linear model for regression

- Simple form of regression
- Picture:

straight line



Problem

- Data: $\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_N}, y_N)\}$
  - $\boldsymbol{x} = < x_1, x_2, \dots, x_M >$: input vector
  - y: target (continuous value)
- Problem: find hypothesis h that maps $\boldsymbol{x}$ to y
  - Assume that h is linear:

$$h(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_M x_M = \boldsymbol{w^T} \binom{1}{\boldsymbol{x}}$$

- Objective: minimize some loss function
  - Euclidean loss: $L_2(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (h(\boldsymbol{x_n}, \boldsymbol{w}) - y_n)^2$

Optimization

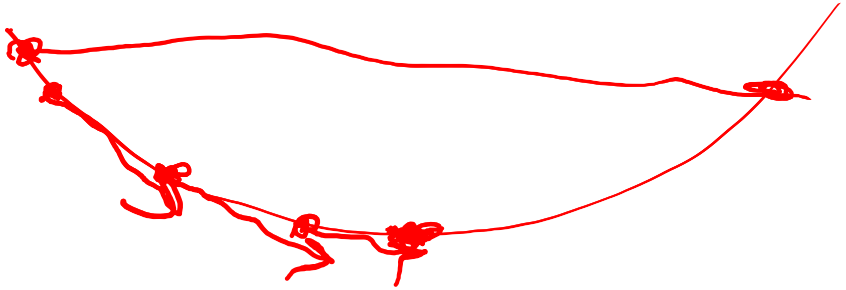
- Find best $\boldsymbol{w}$ that minimizes Euclidean loss

$$\boldsymbol{w}^* = argmin_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^{\boldsymbol{T}} \begin{pmatrix} 1 \\ \boldsymbol{x_n} \end{pmatrix} \right)^2$$

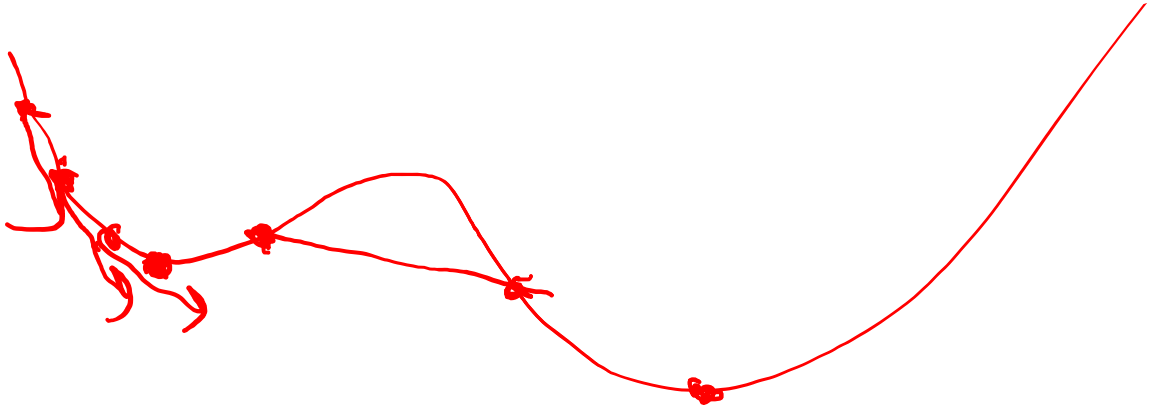
- Convex optimization problem

  $\Longrightarrow$ unique optimum (global)

convex objective



non-convex objective



Solution

- Let $\bar{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$ then $\min_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_{\boldsymbol{n}} \right)^2$
- Find $\boldsymbol{w}^*$ by setting the derivative to 0

$$\frac{\partial L_2}{\partial w_j} = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_{\boldsymbol{n}} \right) \bar{x}_{nj} = 0 \quad \forall j$$

$$\Longrightarrow \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_{\boldsymbol{n}} \right) \bar{\boldsymbol{x}}_{\boldsymbol{n}} = 0$$

- This is a linear system in $\boldsymbol{w}$, therefore we rewrite it as $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$

where $\boldsymbol{A} = \sum_{n=1}^{N} \bar{\boldsymbol{x}}_{\boldsymbol{n}} \bar{\boldsymbol{x}}_{\boldsymbol{n}}^{\boldsymbol{T}}$ and $\boldsymbol{b} = \sum_{n=1}^{N} y_n \bar{\boldsymbol{x}}_{\boldsymbol{n}}$

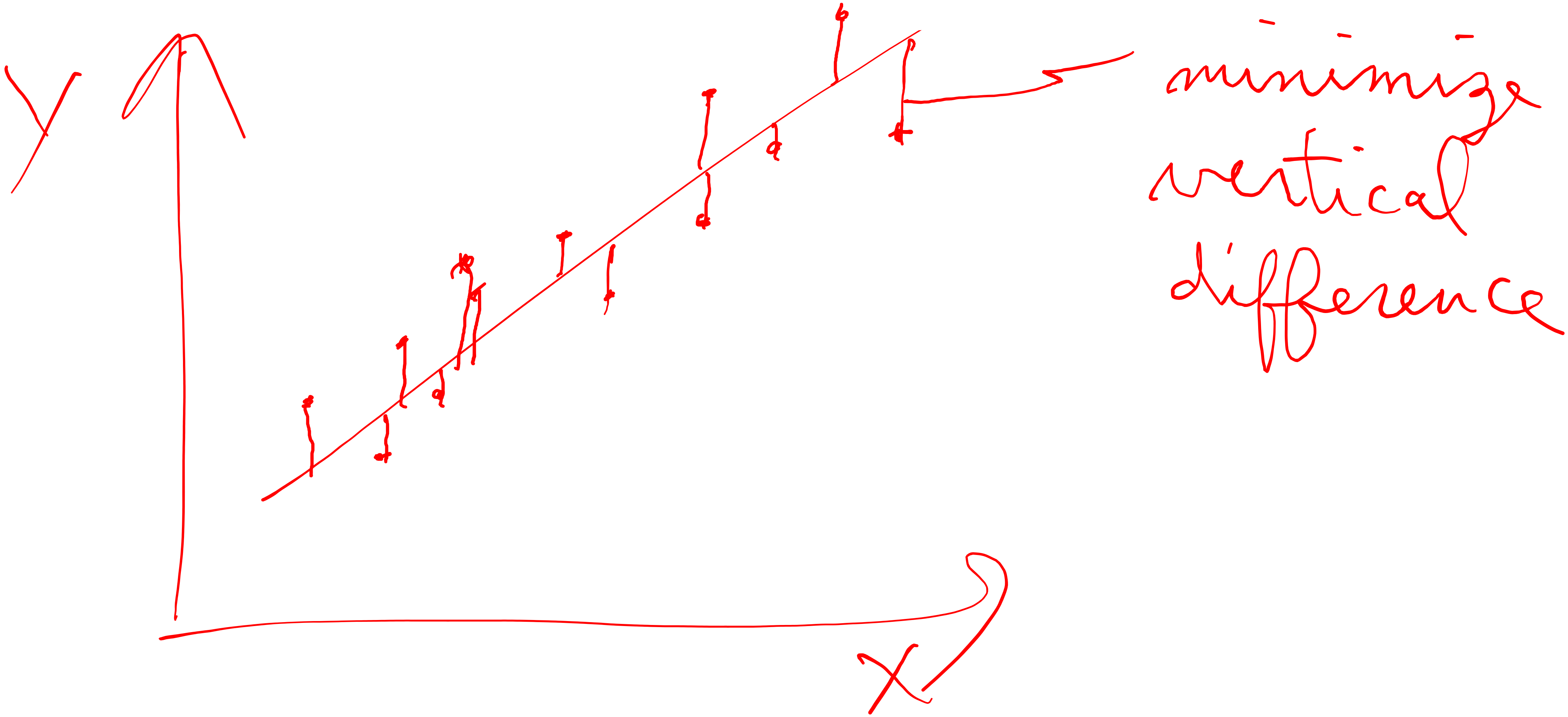
Solution

- If training instances span $\Re^{M+1}$ then $\boldsymbol{A}$ is invertible:

$$\boldsymbol{w} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

- In practice it is faster to solve the linear system $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$ directly instead of inverting $\boldsymbol{A}$
  - Gaussian elimination
  - Conjugate gradient
  - Iterative methods

Picture



Regularization

- Least square solution may not be stable
  - i.e., slight perturbation of the input may cause a dramatic change in the output
  - Form of **overfitting**

Example 1

- Training data: $\bar{\boldsymbol{x}}_1 = \begin{pmatrix}1\\0\end{pmatrix}$ $\quad \bar{\boldsymbol{x}}_2 = \begin{pmatrix}1\\\epsilon\end{pmatrix}$

  $y_1 = 1 \quad y_2 = 1$

- $\boldsymbol{A} = \bar{x}_1\bar{x}_1^T + \bar{x}_2\bar{x}_2^T = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix} + \begin{pmatrix}1\\\varepsilon\end{pmatrix}\begin{pmatrix}1 & \varepsilon\end{pmatrix} = \begin{pmatrix}2 & \varepsilon\\ \varepsilon & \varepsilon^2\end{pmatrix}$

- $\boldsymbol{A}^{-1} = \begin{pmatrix}1 & -1/\varepsilon\\ -1/\varepsilon & 2/\varepsilon^2\end{pmatrix}$ $\quad \boldsymbol{b} = y_1\bar{x}_1 + y_2\bar{x}_2 = 1\begin{pmatrix}1\\0\end{pmatrix} + 1\begin{pmatrix}1\\\varepsilon\end{pmatrix} = \begin{pmatrix}2\\\varepsilon\end{pmatrix}$

- $\boldsymbol{w} = A^{-1}b = \begin{pmatrix}1\\0\end{pmatrix}$

UNIVERSITY OF WATERLOO

Example 2

- Training data: $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$

  $y_1 = 1 + \epsilon$ $y_2 = 1$

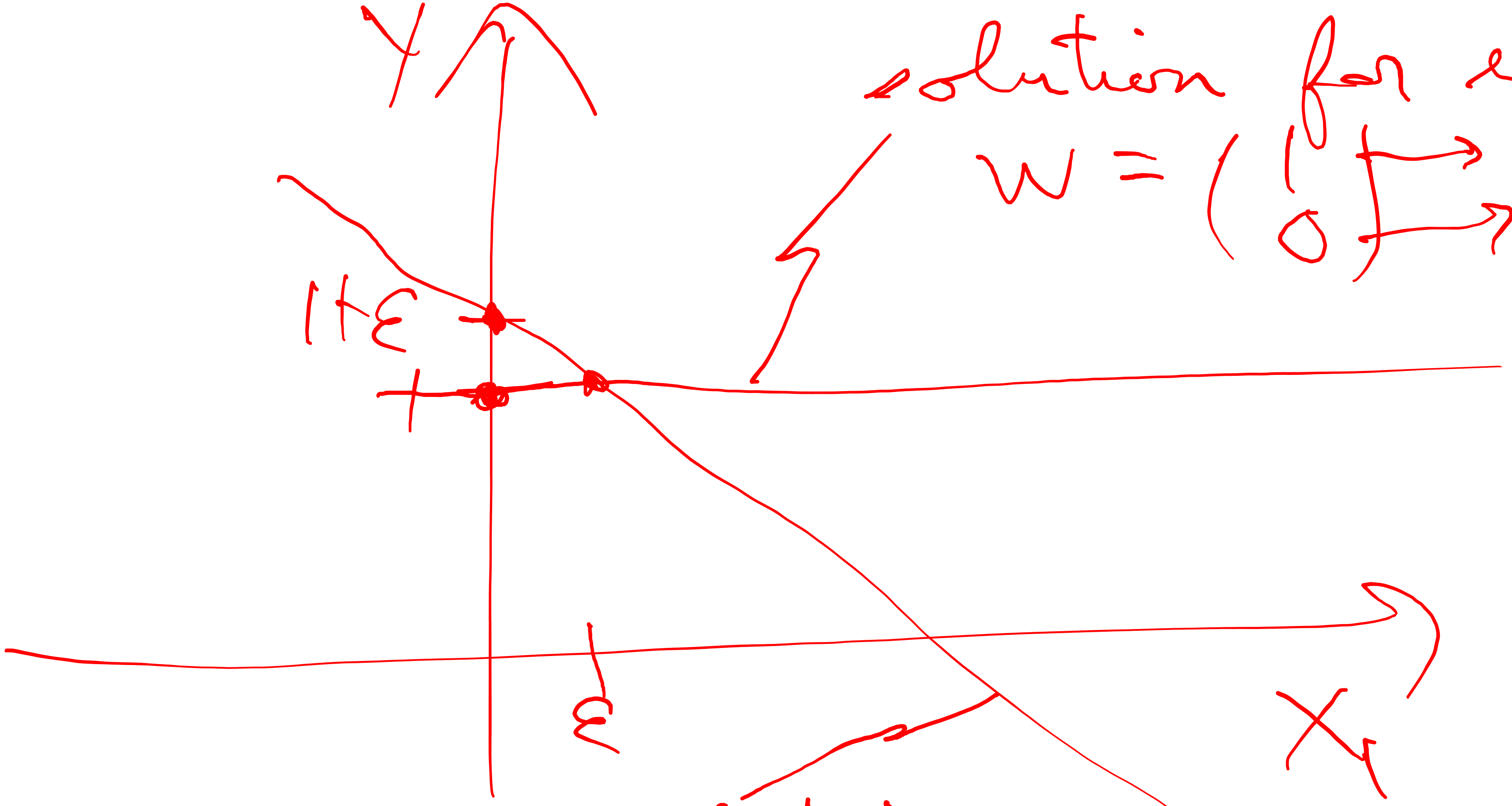
- $\boldsymbol{A} = \begin{pmatrix} 2 & \varepsilon \\ \varepsilon & \varepsilon^2 \end{pmatrix}$

- $\boldsymbol{A}^{-1} = \begin{pmatrix} 1 & -1/\varepsilon \\ -1/\varepsilon & 2/\varepsilon^2 \end{pmatrix}$ $\boldsymbol{b} = y_1 \bar{x}_1 + y_2 \bar{x}_2 = (1+\varepsilon)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 1\begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 2+\varepsilon \\ \varepsilon \end{pmatrix}$

- $\boldsymbol{w} = A^{-1} b = \begin{pmatrix} 1+\varepsilon \\ -1 \end{pmatrix}$

UNIVERSITY OF WATERLOO

Picture



solution for ex. #1

$w = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{array}{l} \rightarrow y \\ \rightarrow \text{slope} \end{array}$

solution for ex. #2

$w = \begin{pmatrix} 1+\varepsilon \\ -1 \end{pmatrix} \begin{array}{l} \rightarrow y \\ \rightarrow \text{slope} \end{array}$

Regularization

norm

square

- Idea: favor smaller values
- Tikhonov regularization: add $||\boldsymbol{w}||_2^2$ as a penalty term

Euclidean norm

- Ridge regression:

$$\boldsymbol{w}^* = argmin_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^{\boldsymbol{T}} \overline{\boldsymbol{x}}_{\boldsymbol{n}}\right)^2 + \frac{\lambda}{2} ||\boldsymbol{w}||_2^2$$

where λ is a weight to adjust the importance of the penalty

Regularization

- Solution: $(\lambda \boldsymbol{I} + \boldsymbol{A})\boldsymbol{w} = \boldsymbol{b}$

identity matrix

- Notes
  - Without regularization: eigenvalues of linear system may be arbitrarily close to 0 and the inverse may have arbitrarily large eigenvalues.
  - With Tikhonov regularization, eigenvalues of linear system are $\geq \lambda$ and therefore bounded away from 0. Similarly, eigenvalues of inverse are bounded above by $1/\lambda$.

Regularized Examples Let $\lambda = 0.05$ & $\varepsilon = 0.1$

Example 1

$$(\lambda I + A) = \begin{pmatrix} 2+\lambda & \varepsilon \\ \varepsilon & \varepsilon^{2}+\lambda \end{pmatrix} = \begin{pmatrix} 2.05 & 0.1 \\ 0.1 & 0.06 \end{pmatrix}$$

$$(\lambda I + A)^{-1} = \begin{pmatrix} 0.531 & -0.885 \\ -0.885 & 18.1416 \end{pmatrix}$$

$$b = \begin{pmatrix} 2 \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 2 \\ 0.1 \end{pmatrix}$$

$$w = \begin{pmatrix} 0.9735 \\ 0.0442 \end{pmatrix}$$

Example 2

$$b = \begin{pmatrix} 2+\varepsilon \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 2.1 \\ 0.1 \end{pmatrix}$$

$$w = \begin{pmatrix} 1.0265 \\ -0.0442 \end{pmatrix}$$