University of Amsterdam
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 3.3 - Supervised Learning
Stochastic Gradient Descent

Erik Bekkers

(Bishop 3.1.3)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Stochastic gradient descent

- for N >> 1 $\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t}$ is very costly to compute!
    - Needs to process all data $(\mathbf{x}_1, ..., \mathbf{x}_N)$ at once.
    - Matrix inversion of M x M matrix: $O(M^3)$
- Loss is a sum of error terms for each datapoint:

$$E_D(\mathbf{w}) = \sum_{i=1}^{N} E(\mathbf{x}_i, t_i, \mathbf{w})$$

$$E(\mathbf{x}_i, t_i, \mathbf{w}) = \frac{1}{2}\left(t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\right)^2$$

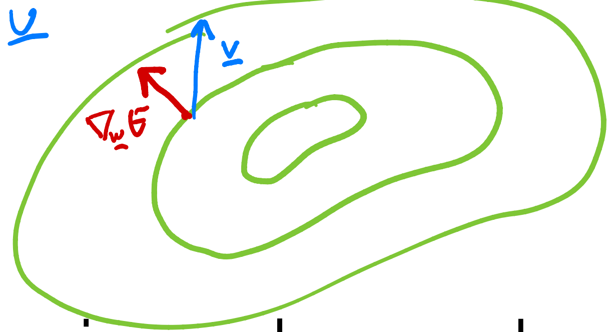
- Approach for large dataset: stochastic gradient descent

approximating the total error with less data points

→ a way for minimizing E_D

Change in error along direction $\underline{v}$

$(\nabla_{\underline{w}} E)\, \underline{v}$



Recap: The Gradient

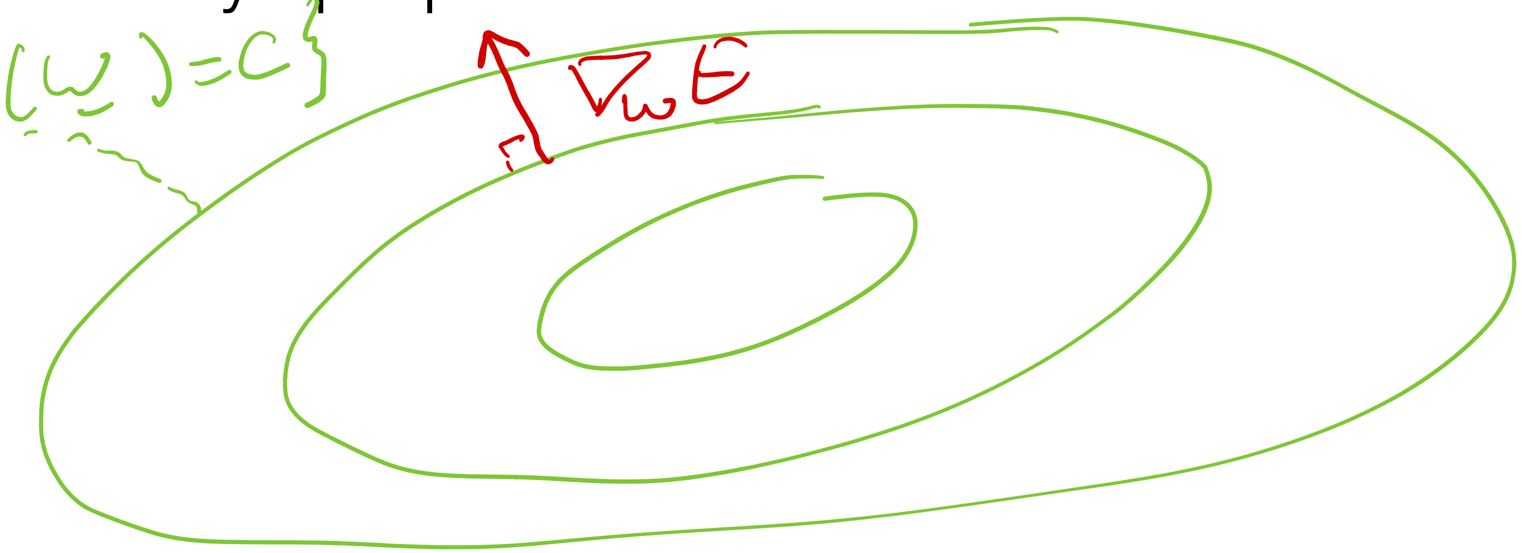
- The gradient encodes all directional derivatives via scalar product

$$\nabla_{\underline{w}} E := \frac{\partial}{\partial \underline{w}} E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_M} \right)$$

$\underline{v} \in \mathbb{R}^M$ directional derivative $(\nabla E)\, \underline{v}$

- The gradient is always perpendicular to the contours of a function

$\{\underline{w} \in \mathbb{R}^m : E(\underline{w}) = c\}$



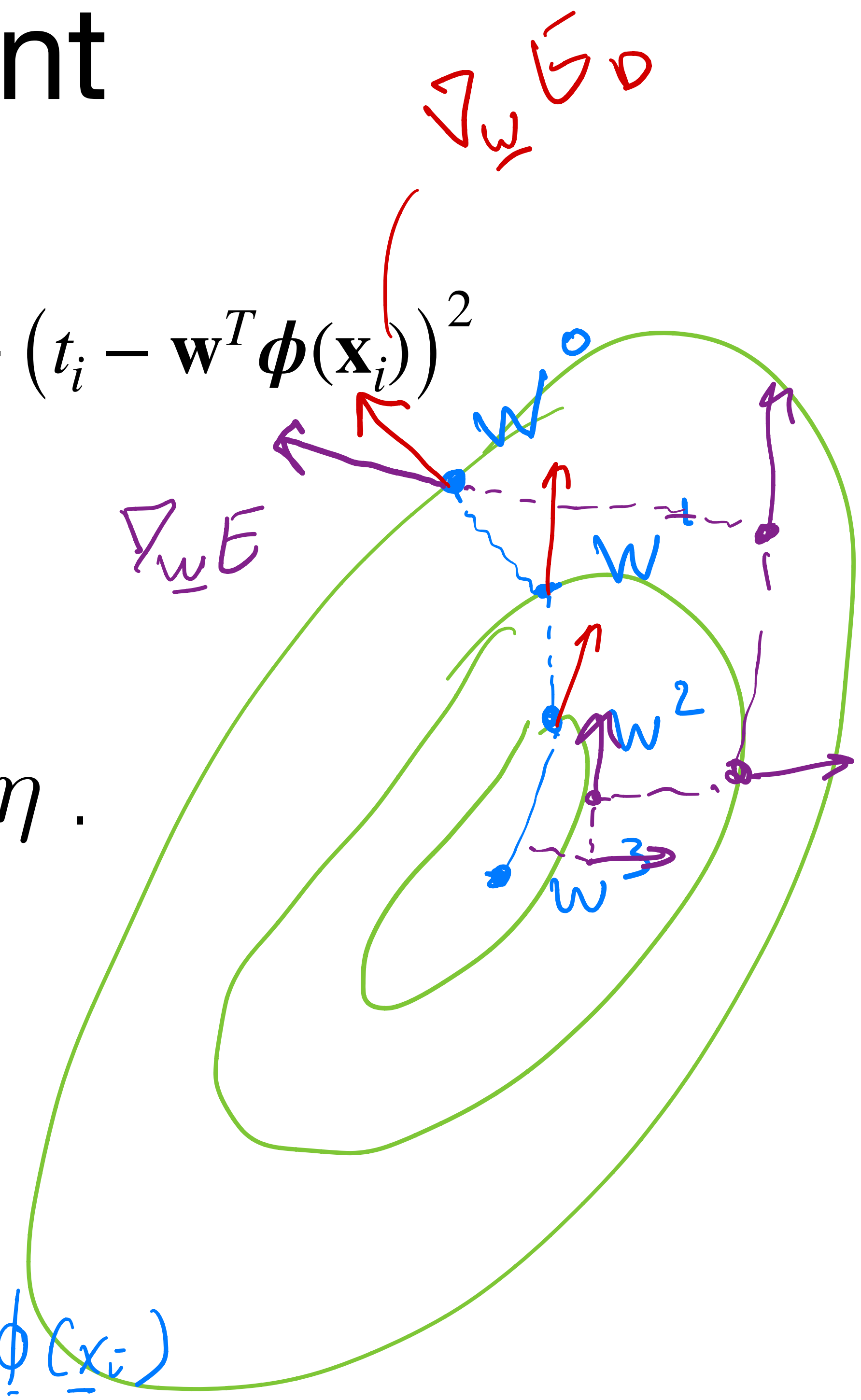
- The gradient always points in the direction of steepest ascent

Bishop

app E

Stochastic gradient descent

$$E_D(\mathbf{w}) = \sum_{i=1}^{N} E(\mathbf{x}_i, t_i, \mathbf{w}) \qquad E(\mathbf{x}_i, t_i, \mathbf{w}) = \frac{1}{2}\left(t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\right)^2$$



- Stochastic gradient descent:
  - Initialize $\mathbf{w}^{(0)}$, choose learning rate η .
  - Iterate over data points, and update

because of our row convention of ∇_w

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta(\nabla_{\mathbf{w}} E(\mathbf{x}_i, t_i, \mathbf{w}^{(\tau)}))^T$$

$$= w^{(\tau)} + \eta(t_i - w^{(\tau)} \phi(x_i))\, \phi(x_i)$$

- If $E_D(\mathbf{w})$ is convex in $\mathbf{w}$ and η is small enough: convergence