Machine Learning
Lecture 7

Agenda

- **Gradient Descent/Ascend (GD)**
  - Concave/Convex functions ($z = f(x,\ y)$)
  - Contourplot/contour lines/level curves ($f(x,\ y) = const$)
  - Gradient (∇f or grad f)
  - Learning rate (η)
- **Stochastic Gradient Descent (SGD)**
  - Epochs
  - Batches
- **Regularization Techniques**
  - Ridged (L_2)
  - Lasso (L_1)
  - Elastic (L_2 and L_1 combined)

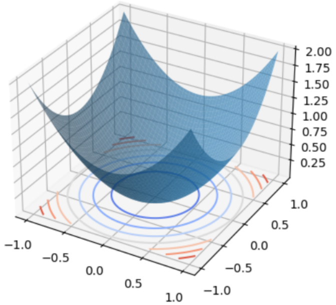
Motivation

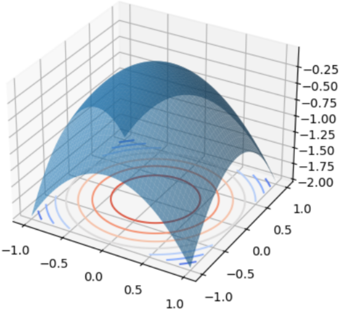
$$a\left(x^{(i)}\right) = w_0 + w_1 \mathrm{x}_1^{(i)} + \ldots + w_d \mathrm{x}_d^{(i)} = w^T \tilde{x},$$

$$w^* = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{y}^T\mathbf{X})^T.$$

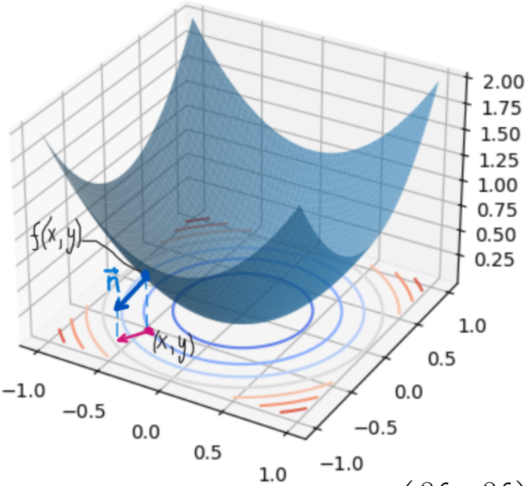
- $\mathbf{X}^T\mathbf{X}$ is (number of features)$\times$(number of features)
- $\exists(\mathbf{X}^T\mathbf{X})^{-1}$?

Convex/Concave Functions



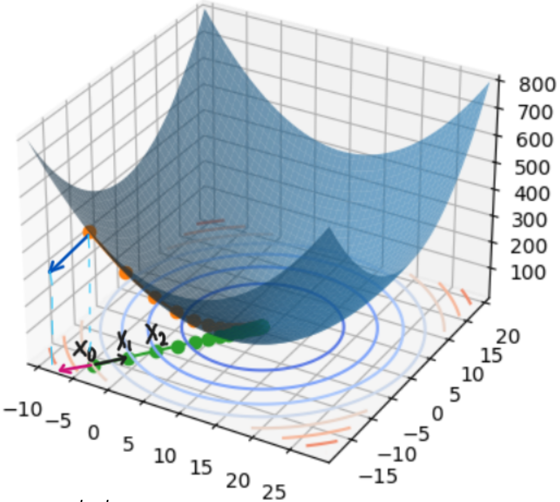


Gradient and Level Curves



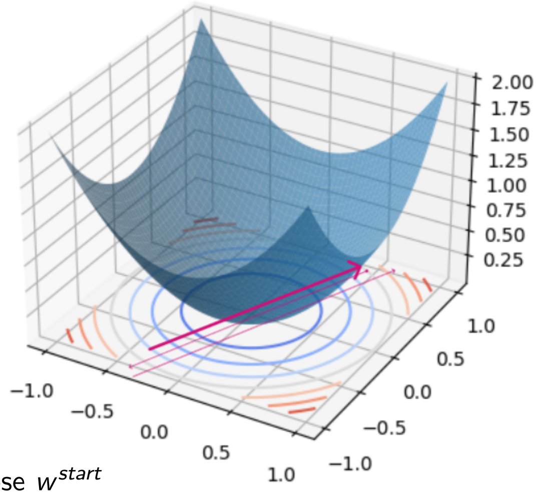
$$\vec{n} = (f_x,\ f_y,\ -1) \qquad \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

(Batch) Gradient Descent



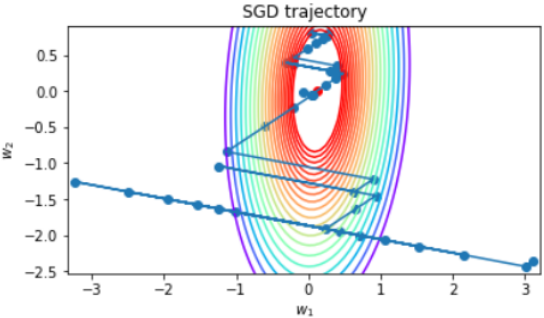
1. Choose w^{start}
2. $w^{new} = w^{old} - \nabla Loss\left(\mathbf{X} w^{old},\ \mathbf{y}\right)$,
3. Stop after M iterations or $\left|w^{new} - w^{old}\right| < \varepsilon$.

Learning Rate



1. Choose w^{start}
2. $w^{new} = w^{old} - \eta \nabla Loss\left(\mathbf{X} w^{old},\ \mathbf{y}\right)$,
3. Stop after M iterations or $\left|w^{new} - w^{old}\right| < \varepsilon$.

Stochastic Gradient Descent



1. Choose w^{start}
2. $w^{new} = w^{old} - \eta \nabla Loss\left((w^{old})^T \tilde{x}^{(i)},\ y^{(i)}\right)$,
3. Stop after $N \times M$ *epoch* iterations or $|w^{new} - w^{old}| < \varepsilon$.

Mini-Batch SGD

Batch $\mathbf{X}_1$	$x^{(1)}$	$y^{(1)}$
	$x^{(2)}$	$y^{(2)}$
	...	...
	$x^{(N_1)}$	$y^{(N_1)}$
Batch $\mathbf{X}_2$	$x^{(N_1+1)}$	$y^{(N_1+1)}$
	$x^{(N_1+2)}$	$y^{(N_1+2)}$
	...	...
	$x^{(2N_1)}$	$y^{(2N_1)}$
Batch $\mathbf{X}_B$	$x^{((B-1)N_1+1)}$	$y^{((B-1)N_1+1)}$
	$x^{((B-1)N_1+2)}$	$y^{((B-1)N_1+2)}$
	...	...
	$x^{(N_B)}$	$y^{(N_B)}$

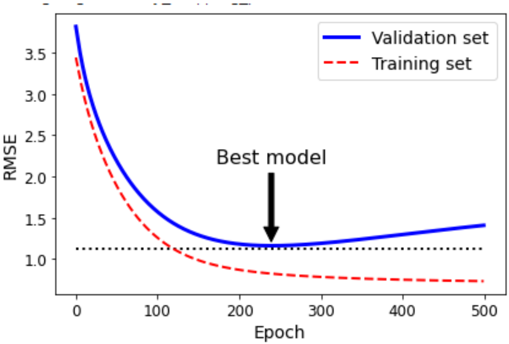
SGD

$$Loss_{batch} = \frac{1}{N} \sum_{1}^{N} \left(a\left(x^{i}\right) - y^{(i)} \right)^{2}$$

$$Loss_{SGD} = \left(a\left(x^{i}\right) - y^{(i)} \right)^{2}$$

$$Loss_{mini-batch} = \frac{1}{N_1} \sum_{(b-1)N_1+1}^{bN_1} \left(a\left(x^{i}\right) - y^{(i)} \right)^{2}$$

Early Stopping



https://github.com/ageron/handson-ml2/blob/master/04_training_linear_models.ipynb

Regularization

Ridged (L_2):

$$Loss + \alpha \|w_{-0}\|_2^2 \rightarrow min$$

Lasso (L_1)

$$Loss + \beta \|w_{-0}\|_1 \rightarrow min$$

Elastic (L_2 and L_1 combined)

$$Loss + \alpha \|w_{-0}\|_2 + \beta \|w_{-0}\|_1 \rightarrow min$$

Usually, w_0 is not included, i.e., $\|w_{-0}\|_1 = |w_1| + \ldots + |w_d|$ and $\|w_{-0}\|_2^2 = w_1^2 + \ldots + w_d^2$.