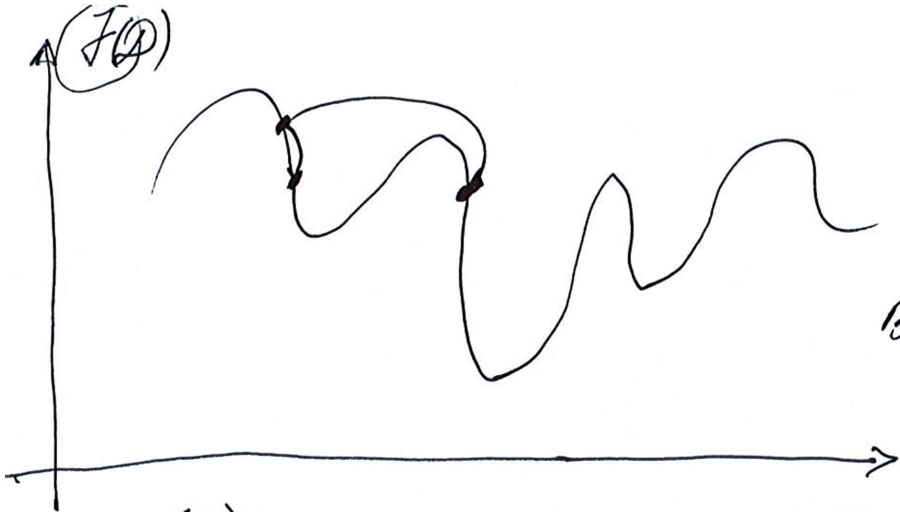
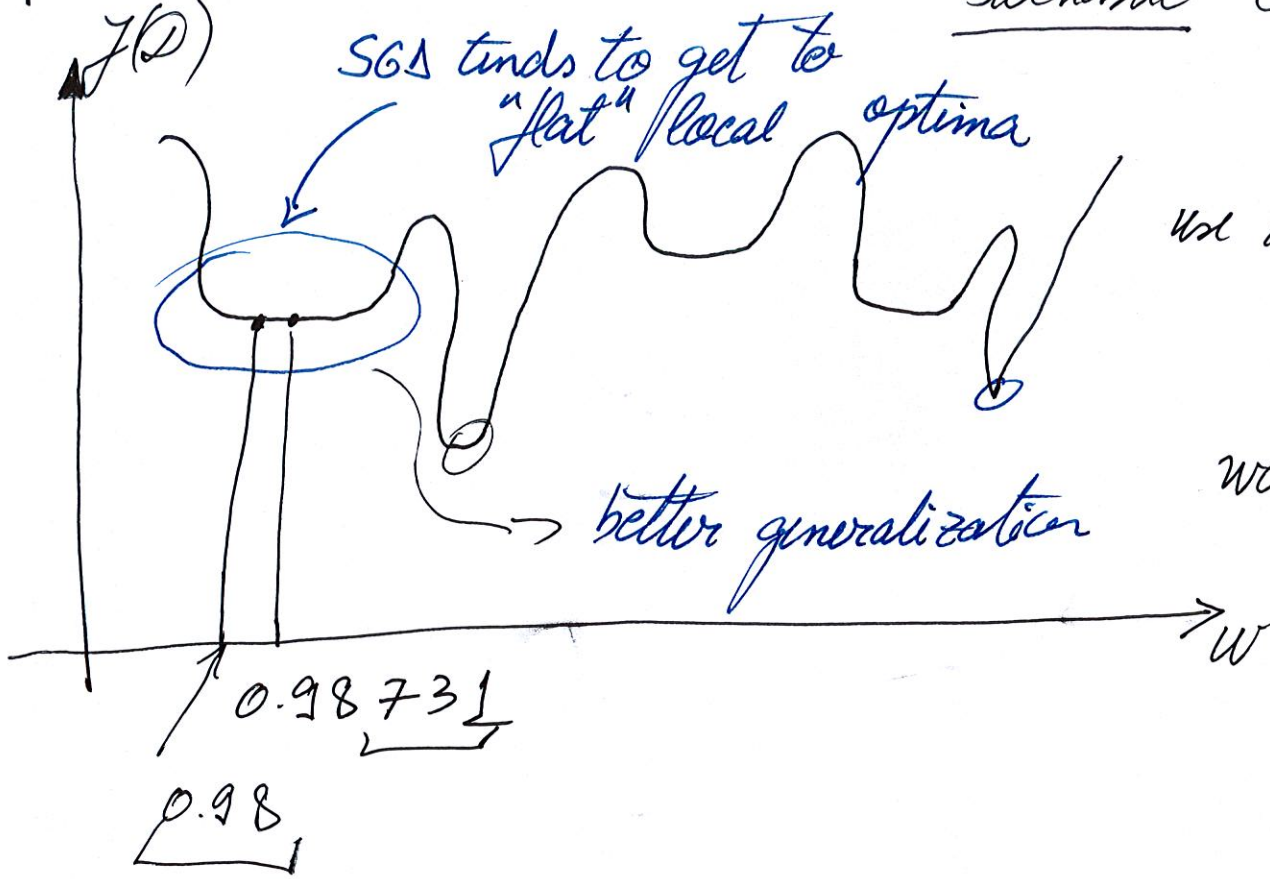


<u>Stochastic</u> w/ mini-batch B of entire dataset D, $B \subseteq D$

Batch GD: Compute $J(D)$ and use $\frac{\partial J(D)}{\partial w}$

<u>Stochastic</u>: Compute $J(B)$, use $\frac{\partial J(B)}{\partial w}$

use it as a substitute for $\frac{\partial J(D)}{\partial w}$, want to minimize $J(D)$.



Linear regression

Batch GD

$$J(w) = \frac{1}{2N} \sum_{n=1}^{N} \left(w^T x^{(n)} - t_n\right)^2$$

$$\frac{\partial J}{\partial w} = \frac{1}{N} \sum_{n=1}^{N} \left(w^T x^{(n)} - t_n\right) \cdot x^{(n)}$$

$$w^{t+1} = w^t - \eta \frac{\partial J}{\partial w}$$

Stochastic GD (minibatch size 1).

$$J(w) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\frac{\left(w^T x^{(n)} - t_n\right)^2}{2}}_{\to J_n(w)}$$

$$\frac{\partial J_n}{\partial w} = \left(w^T x^{(n)} - t_n\right) x^{(n)}$$

$$w^{t+1} = w^t - \eta \frac{\partial J_n}{\partial w}$$

loop over $n = 1 \ldots N$, $e = 1 \ldots E$ or convergence.

LMS (least mean-square) algorithm.

[0,1] scaling:

$$x_j \longrightarrow \frac{x_j - \min_j}{\max_j - \min_j} = \hat{x}_j$$

$$\min_j = \min_{1 \le n \le N} x_j^{(n)}$$

$$\max_j = \max_{1 \le n \le N} x_j^{(n)}$$

Linear transform.

$$\hat{x}_j = \frac{1}{\max_j - \min_j} \cdot x_j - \frac{\min_j}{\max_j - \min_j}$$

n	x_j	$\hat{x}_j$	
1	2	0.2	Train
2	1	0	
3	6	1	
4	8	?	Test

Scaling to [0,1]

$\min_j = 1 \quad \max_j = 6$

$n=1: \quad \hat{x}_j = \dfrac{x_j - \min_j}{\max_j - \min_j} = \dfrac{2-1}{6-1} = \dfrac{1}{5}$

$n=2: \quad \hat{x}_j = \dfrac{1-1}{6-1} = 0$

$n=3: \quad \hat{x}_j = \dfrac{6-1}{6-1} = 1$

$n=4 \quad \hat{x}_j = \dfrac{8-1}{6-1} = \dfrac{7}{5} = 1.4$ ①

clipping: clip to [0, 1] ②

$\hat{x}_j = 1.$

Standardization

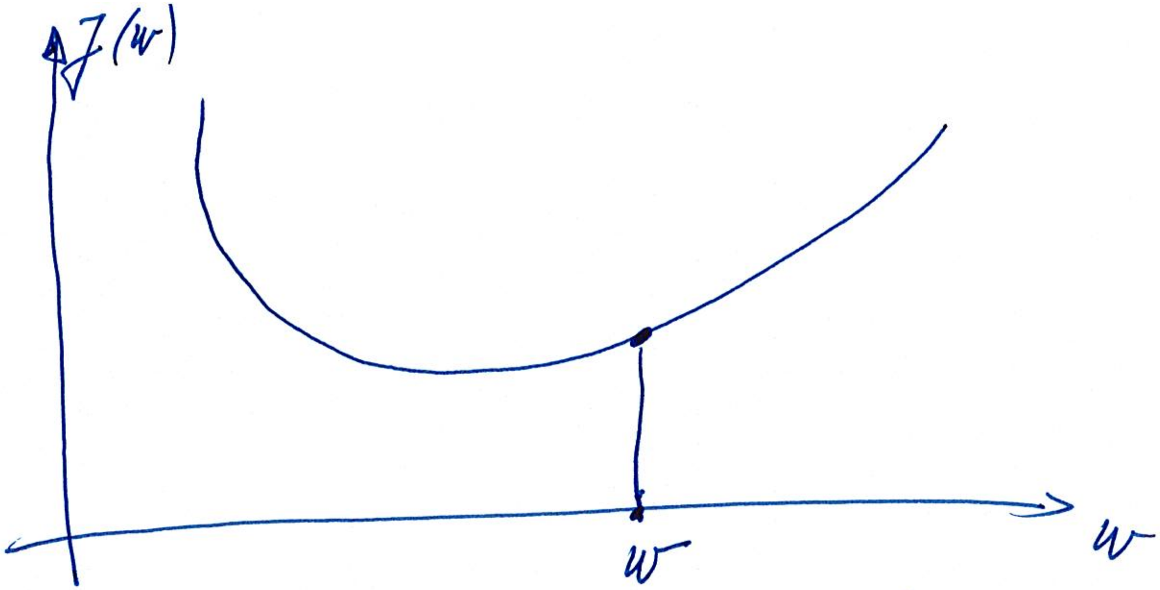
- sample mean μ_j:

$$\mu_j = \frac{1}{N}\sum_{n=1}^{N} x_j^{(n)}$$

- sample std.dev σ_j:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(x_j^{(n)} - \mu_j\right)^2}$$

- $\hat{x}_j = \dfrac{x_j - \mu_j}{\sigma_j}$



$$J'(w) = \frac{\partial J}{\partial w} = \lim_{\varepsilon \to 0} \frac{J(w+\varepsilon) - J(w-\varepsilon)}{2\varepsilon}$$

Set $\varepsilon = 0.001$, compute $\underline{J'_{num}(w_j)} = \dfrac{J(w_j+\varepsilon) - J(w_j-\varepsilon)}{2\varepsilon}$

$\underline{J_{grad}}$ = ... gradient( )

$$\left| \frac{J'_{num}(w_j) - J_{grad}[j]}{J_{grad}[j]} \right| < 10^{-8}$$