Stochastic Gradient Descent updates weights using part of the data

$$\underbrace{f(w)}_{} = \underbrace{l(w)}_{\text{"loss"}} + \lambda \underbrace{r(w)}_{\text{"regularizer"}} \qquad w^{(k+1)} = w^{(k)} - \frac{\tau}{2} \underbrace{\nabla_w f(w)}_{\text{gradient}}$$

$$l(w) = \underbrace{\sum_{i=1}^{N} (d_i - x_i^T w)^2}_{\text{squared error}} \qquad l(w) = \underbrace{\sum_{i=1}^{N} (1 - d_i x_i^T w)_+}_{\text{hinge loss}} \qquad (\underset{\text{labels}}{d_i}, \underset{\text{features}}{x_i}),\ i = 1, \cdots, N$$

$$\underbrace{\nabla_w l(w) = -2 \sum_{i=1}^{N} (d_i - x_i^T w) x_i \qquad \nabla_w l(w) = -\sum_{i=1}^{N} I\{d_i x_i^T w < 1\} x_i}_{\text{depends on all data}}$$

SGD: $f(w) = \sum_{i=1}^{N} f_i(w)$ Define i_k, $k = 1, 2, \cdots$

$w^{(k+1)} = w^{(k)} - \frac{\tau}{2} \nabla_w f_{i_k}(w^{(k)})$ depends on one sample (d_{i_k}, x_{i_k})

SGD cycles through training data

1.) Cyclical (incremental gradient descent)

$i_k = k \bmod N$, e.g. $i_k = 1, 2, 3, 4, 1, 2, 3, 4, \ldots$

2). Random permutation (reshuffle every N rounds)

$i_k = \underbrace{2,4,1,3}, \underbrace{2,1,4,3}, \underbrace{4,3,1,2}$

3). Stochastic Gradient Descent (uniformly at random)

$i_k = \text{uniform}\{1, 2, \cdots, N\}$ $\quad i_k = 2,1,3,1,4,4,2,3,1$

update by $-\frac{\tau}{2} \nabla_w f_{i_k}(w)$ at each iteration

on average gives gradient $E\{\nabla_w f_{i_k}(w)\} \approx \frac{\nabla_w f(w)}{N}$

SGD has computational benefits

1. Computing $\nabla_w f_{i_k}(w^{(k)})$ is easier / faster than $\nabla_w f(w^{(k)})$

2. May not be able to store x_i, $i = 1 \cdots N$ in memory

3. Noisy gradient $\nabla_w f_{i_k}(w^{(k)})$ introduces added regularization

Example : Ridge Regression

$$f(w) = \sum_{i=1}^{N} (d_i - x_i^T w)^T + \lambda \|w\|_2^2 = \sum_{i=1}^{N} \underbrace{\left\{ (d_i - x_i^T w)^T + \frac{\lambda}{N} \|w\|_2^2 \right\}}_{f_i(w)}$$

$$\nabla_w f_i(w) = \nabla_w \left[(d_i - x_i^T w)^2 + \frac{\lambda}{N} w^T w \right]$$

$$= -2 (d_i - x_i^T w) x_i^T + 2 \frac{\lambda}{N} w$$

$$w^{(k+1)} = w^{(k)} - \frac{\tau}{2} \nabla_{w^{(k)}} f_{i_k}(w^{(k)})$$

$$= w^{(k)} + \tau (d_{i_k} - x_{i_k}^T w^{(k)}) x_{i_k} - \frac{\tau \lambda}{N} w^{(k)}$$

vs. $w^{(k+1)} = w^{(k)} + \tau A^T (A w^{(k)} - d) - \lambda \tau w^{(k)}$ $\quad A: N \times M$