Stochastic Gradient Descent

Objectives

- Simplify gradient descent update
- Common methods for cycling through data
- Benefits
- Examples

Stochastic gradient descent updates weights using part of the data

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

squared error

$$l(\underline{w}) = \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})^2$$

hinge loss

$$l(\underline{w}) = \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+$$

$(d_i, \underline{x}_i), i = 1, \ldots, N$

labels features

$$\nabla_{\underline{w}} l(\underline{w}) = -2 \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w}) \underline{x}_i \qquad \nabla_{\underline{w}} l(\underline{w}) = -\sum_{i=1}^{N} I_{\{d_i \underline{x}_i^T \underline{w} < 1\}} \underline{x}_i$$

depends on all the data

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

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

depends on one sample $(d_{i_k}, \underline{x}_{i_k})$

SGD cycles through training data

1) Cyclical (incremental gradient descent)

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

2) Random permutation (reshuffle every N rounds)

$i_k = 2,4,1,3,\boxed{2,1,4,3,}\ 4,3,1,2 \ \ldots$

3) Stochastic gradient descent (uniformly at random)

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

Update by $-\frac{\tau}{2}\nabla_{\underline{w}} f_{i_k}(\underline{w})$ at each iteration

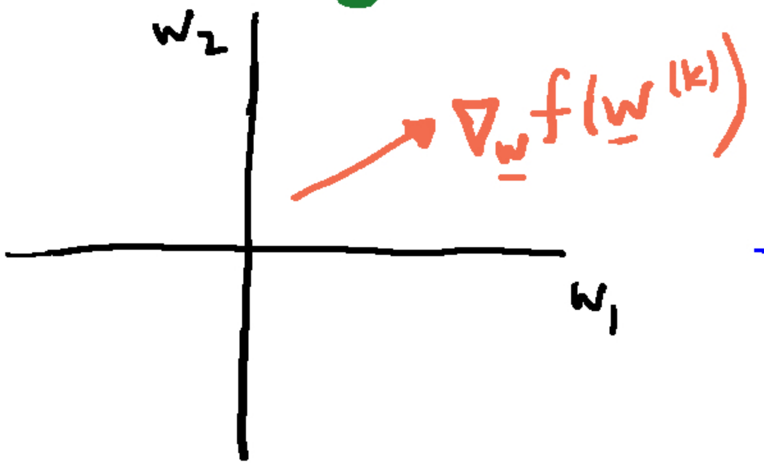
On average gives gradient $E\{\nabla_{\underline{w}} f_{i_k}(\underline{w})\} \simeq \frac{\nabla_{\underline{w}} f(\underline{w})}{N}$

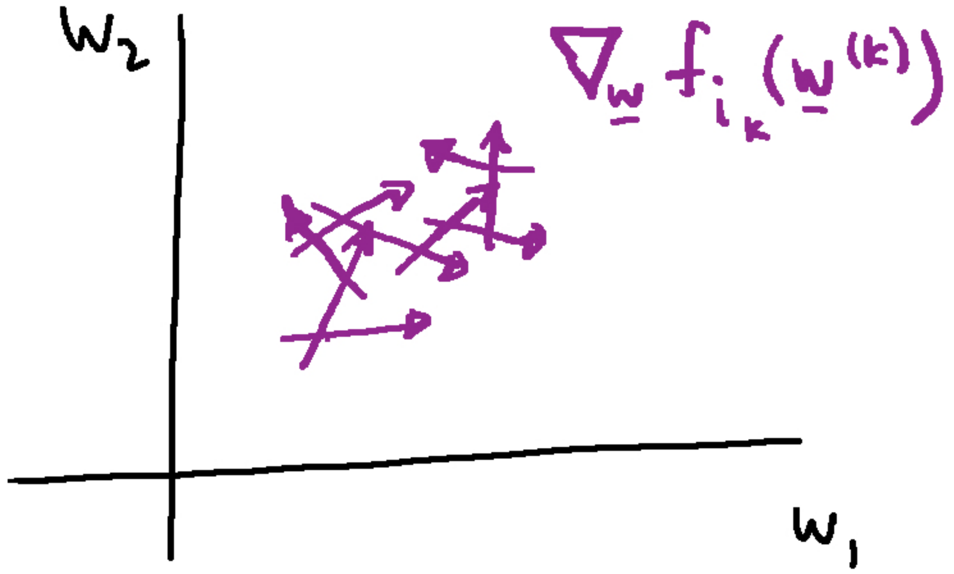
SGD has computational benefits

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

2) May not be able to store $\underline{x}_i,\ i=1,\dots,N$ in memory

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





Example: Ridge Regression

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

$$\nabla_{\underline{w}} f_i(\underline{w}) = \nabla_{\underline{w}} \left[(d_i - \underline{x}_i^T \underline{w})^2 + \frac{\lambda}{N} \underline{w}^T \underline{w} \right]$$

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

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

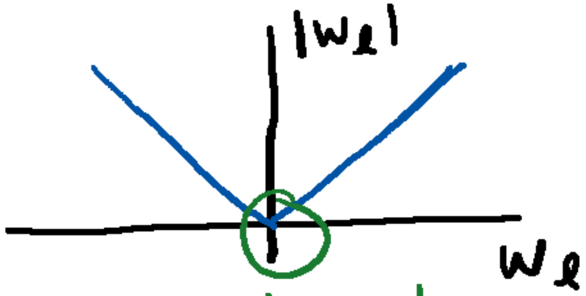
$$= \underline{w}^{(k)} + \tau (d_{i_k} - \underline{x}_{i_k}^T \underline{w}^{(k)}) \underline{x}_{i_k} - \frac{\tau\lambda}{N} \underline{w}^{(k)}$$

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

Example: Gradient descent for LASSO

$$f(\underline{w}) = \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})^2 + \lambda \|\underline{w}\|_1 = \sum_{i=1}^{N} \underbrace{\left\{ (d_i - \underline{x}_i^T \underline{w})^2 + \frac{\lambda}{N} \|\underline{w}\|_1 \right\}}_{f_i(\underline{w})}$$

Consider $\nabla_{\underline{w}} \sum_{\ell=1}^{M} |w_\ell|$



use subgradient

$$\frac{d}{dw_\ell}|w_\ell| = \begin{cases} \text{sign}(w_\ell) & w_\ell \neq 0 \\ [-1,1] & w_\ell = 0 \end{cases}$$

"0" popular

Write $\nabla_{\underline{w}} \|\underline{w}\|_1 = \text{sign}(\underline{w})$

$$\nabla_w f_i(\underline{w}) = -2(d_i - \underline{x}_i^T \underline{w})\underline{x}_i + \frac{\lambda}{N} \text{sign}(\underline{w})$$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} + \tau \left(d_{i_k} - \underline{x}_{i_k}^T \underline{w}^{(k)} \right) \underline{x}_{i_k} - \frac{\lambda \tau}{2N} \text{sign}(\underline{w}^{(k)})$$