CS/ECE/ME 532 Period 20

- Unit 5 Quiz Thursday next week

- Today
  - Subgradients
  - Stochastic Gradient Descent (SGD)

Applications of SGD –

- any ML problem that involves lots of training data, for example:
  - image/video classification and recognition
  - ML translation
  - large scale prediction and regression tasks

Stochastic Gradient Descent and Sub-gradients

Classifying new data:

$\boldsymbol{x} =$ 

$\widehat{y} = \text{sign}(\boldsymbol{x}^T \boldsymbol{w})$

if $\widehat{y} = -1$ then cat

if $\widehat{y} = 1$ then dog

Training a classifier:

$$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}) \qquad \min_{\boldsymbol{w}} \overbrace{\sum_{i=1}}^{\text{a million}} (\boldsymbol{x}_i^T \boldsymbol{w} - y_i)^2$$

$(\boldsymbol{x}_i, y_i), i = 1, ..., \text{a million}$

Problem: could be PB of training data

<u>SGD Main idea:</u>

Do gradient descent,
but on a random subset of training examples at each iteration.

Parallelize training!

Stochastic Gradient Descent

$$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}) \qquad \boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \tau \nabla \ell(\boldsymbol{w}^k)$$

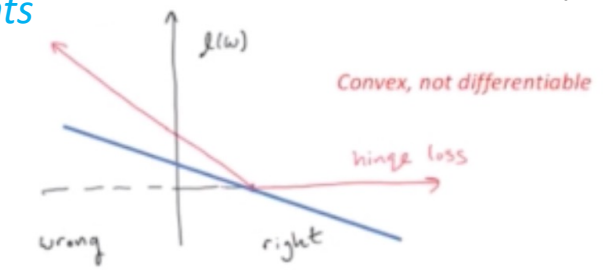
single training example $(\boldsymbol{x}_i, y_i)$

$$\min_{\boldsymbol{w}} \overbrace{\sum_{i=1}}^{\text{a million}} \ell_i(\boldsymbol{w})$$

$$\boldsymbol{w}^{(1)} = \boldsymbol{w}^{(0)} - \tau \sum_{i=1}^{100} \nabla \ell_i(\boldsymbol{w}^{(0)})$$

$$\boldsymbol{w}^{(2)} = \boldsymbol{w}^{(1)} - \tau \sum_{i=101}^{200} \nabla \ell_i(\boldsymbol{w}^{(1)})$$

Subgradients



Subgradient – any plane that lies below function.

subgradient: any $\boldsymbol{v}$ such that

$\ell(\boldsymbol{w}) \geq \ell(\boldsymbol{w}_0) + (\boldsymbol{w} - \boldsymbol{w}_0)^T \boldsymbol{v}$