Backpropagation & SGD

parameters

$h(x) = w^T \phi(x)$

$\phi(x) = \sigma(A \phi'(x))$ — $a(x)$

$\phi'(x) = \sigma(A' \phi''(x))$ — $a'(x)$

$\phi''(x) = \sigma(A'' x)$ — $a''(x)$

transition function

$$\mathcal{L}(h) = \sum_{i=1}^{n} \ell(h(x_i), y_i) = \sum_{i=1}^{n} (h(x_i) - y_i)^2$$

$$\frac{\partial \mathcal{L}}{\partial w} = \sum_{i=1}^{n} (w^T \phi(x_i) - y_i) \phi(x_i)$$

Backpropagation

$$\frac{\partial \mathcal{L}}{\partial A} = \frac{\partial \mathcal{L}}{\partial a} \frac{\partial a}{\partial A} \qquad a(x) = A\phi'(x)$$

$$\frac{\partial \mathcal{L}}{\partial A'} = \frac{\partial \mathcal{L}}{\partial a} \frac{\partial a}{\partial a'} \frac{\partial a'}{\partial A'}$$

$$w \leftarrow w - \alpha \frac{\partial \mathcal{L}}{\partial w}$$

$$A \leftarrow A - \alpha \frac{\partial \mathcal{L}}{\partial A}$$

Convex Optimization (SVMs, Ridge Regression, ...)

$\phi'(x) = \sigma(a')$ $\quad \frac{\partial a}{\partial a'} = A \frac{\partial \sigma(a')}{\partial a'}$

$a = A\sigma(a')$

ReLU(x) $\max\{0, x\}$

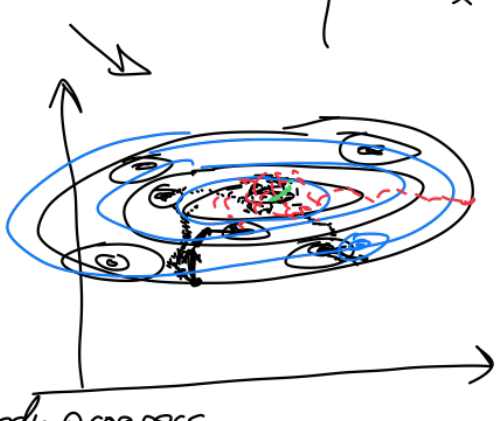
Non-convex Optimization (NNs)

Normal GD

$$\frac{\partial \mathcal{L}}{\partial A} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial A}$$

SGD

$$\frac{\partial \mathcal{L}}{\partial A} \approx \frac{\partial \ell_i}{\partial A}$$



Important Changes

① Initialization: random!

② SGD

1. Small but steady progress
2. Wide and deep minima