Lecture 6

Optimization continued

Gradient Descent (Recap)

Objective fn: $f(\theta)$

Init θ at θ_0

For $t=1 \dots T$

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$$

ML: $f(\theta)$: training loss

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_\theta(x^{(i)}), y^{(i)})$$

$$\nabla f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(h_\theta(x^{(i)}, y^{(i)})$$

$|B| < n$

Stochastic Gradient Descent

θ : init at θ_0

For $t = 1 \ldots T$

sample some indices B

$$\theta_{t+1} = \theta_t - \eta \frac{1}{|B|} \sum_{i \in B} \nabla f(h_\theta(x^{(i)}, y^{(i)})$$

each iteration is cheaper

$\nabla f(\theta)$ Empirical train loss gradient

$\searrow$ on just B samples "Stochastic gradient"

Stochastic gradient descent $\equiv$ tolerate some noise in gradient at each step

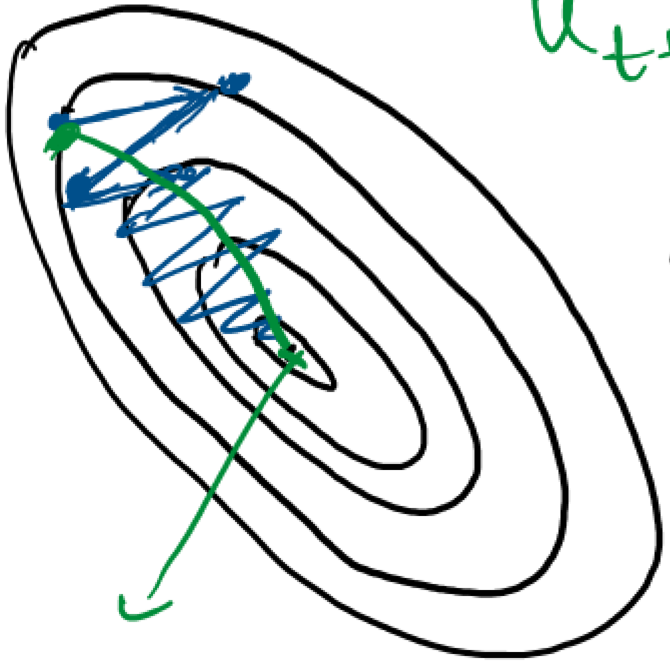
☼ Todo reference for SGD vs GD

SGD + Momentum

"level sets" of a function

$$\boxed{u_{t+1} = \beta u_t + \nabla f(\theta_t)} \text{ equiv}$$

$$u_{t+1} = \beta u_t + (1-\beta)\nabla f(\theta_t)$$



$$\theta_{t+1} = \theta_t - \eta\, u_{t+1}$$

overshoot at the minimum

SGD + "adaptivity"

(f_θ): $\theta = [$all the weights $\dots]$

$$\nabla_\theta f(\theta) = [2 \quad 100 \quad -0.0005 \quad \dots \dots]$$

same across all weights

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta f(\theta)$$

diff ~~f~~weights updated wildly differently

RMS Prop

Normalize based on magnitude of gradients

$$s_{t+1} = \gamma s_t + (1-\gamma)\left[\nabla f(\theta_t)\right]^2 \rightarrow \text{elem-wise}$$

"norm" or magnitude

$$\begin{bmatrix} 1 & 0.01 & 100 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 10^{-4} & 10^4 \end{bmatrix}$$

$$\theta_{t+1} = \theta_t - \frac{\eta \nabla f(\theta_t)}{\sqrt{s_{t+1}} + \epsilon}$$

elem wise

prevent dividing by zero

Adam

Adam : $\underbrace{\text{adaptive}}_{\text{RMS prop}} + \underbrace{\text{momentum}}_{\text{SGD w/ momentum}}$

Bias - correction

$$u_{t+1} = \beta u_t + (1-\beta) \nabla f(\theta_t)$$

$$\hat{u}_{t+1} = \frac{u_{t+1}}{1-\beta^{t+1}}$$

→ β to the power t

$$\theta = \theta - \eta \hat{u}_{t+1}$$

$$u_1 = (1-\beta)\, g_1$$

$$u_2 = \beta u_1 + (1-\beta)\, g_2$$

$$= \beta(1-\beta)\, g_1 + (1-\beta)\, g_2$$

$$\beta(1-\beta) + (1-\beta) \approx 1$$

$$u_2 : (1-\beta)^2 \text{ as } dr$$

moving avg does not have bias correction momentum : $\hat{u}_{t+1} = \dfrac{u_{t+1}}{(1-\beta^{t+1})}$

adaptivity : $\hat{s}_{t+1} = \dfrac{s_{t+1}}{(1-\gamma^{t+1})}$

$$\boxed{\theta_{t+1} = \theta_t - \frac{\eta\, \hat{u}_{t+1}}{\sqrt{\hat{s}_{t+1}} + \epsilon}}$$

bias-corrected momentum

bias-corrected adaptivity or RMSprop

Optimization vs generalization

minimizing train loss → how we do this?
Adam, SGD ...

→ SGD is good (better than GD)
at test loss

noise acts as a regularizer

Adam vs SGD

often faster at optimizing "overfitting"
but does worse on test

Optimization is important but not the only thing to care about. Optimization is about small train loss but we care about good $\boxed{\text{test loss}}$