Optimization Methods for Deep Learning

Gradient descent for non-convex optimization

$\min_x f(x)$

Descent Lemma: Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable, and $\|\nabla^2 f\|_2 \leq \beta$. Then setting the learning rate $\eta = 1/\beta$, and applying gradient descent, $x_{t+1} = x_t - \eta \nabla f(x_t)$, we have:

$$f(x_t) - f(x_{t+1}) \geq \frac{1}{2\beta} \|\nabla f(x_t)\|_2^2.$$

Pf: by Taylor expansion & mean-value theorem

$$f(x+\delta) = f(x) + \delta^T \nabla f(x) + \frac{1}{2} \delta^T \nabla^2 f(y) \delta \quad \text{for some } y$$

$$\delta^T \nabla^2 f(y) \delta \leq \|\nabla^2 f(y)\|_2 \cdot \|\delta\|_2^2 \leq \beta \cdot \|\delta\|_2^2$$

choose $\delta = -\eta \nabla f(x_t)$

$$f(x_{t+1}) \leq f(x_t) - \eta \|\nabla f(x_t)\|_2^2 + \frac{1}{2}\eta^2 \beta \|\nabla f(x_t)\|_2^2$$

$$\leq f(x_t) - \frac{1}{2}\eta \|\nabla f(x_t)\|_2^2$$

if $\eta = \frac{1}{\beta}$

$$= f(x_t) - \frac{1}{2\beta} \|\nabla f(x_t)\|_2^2$$

Converging to stationary points

Theorem: In $T = O(\frac{\beta}{\epsilon^2})$ iterations, we have $\|\nabla f(x)\|_2 \leq \epsilon$.

Pf: $f(x_{t+1}) \leq f(x_t) - \frac{1/\beta}{2} \|\nabla f(x_t)\|_2^2$

sum over $t = 0, \ldots, T-1$

$$\sum_{t=1}^{T} f(x_t) \leq \sum_{t=0}^{T-1} f(x_t) - \frac{1/\beta}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad f(x_T) \leq f(x_0) - \frac{1/\beta}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad \underbrace{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2}_{T \cdot \min_{0 \leq t \leq T-1} \|\nabla f(x_t)\|_2^2 \leq} \leq 2 \cdot \frac{f(x_0) - f(x_T)}{1/\beta}$$

$$\Rightarrow \quad \min_{0 \leq t \leq T-1} \|\nabla f(x_t)\| \leq \sqrt{\frac{2(f(x_0) - f(x_T))}{1/\beta \cdot T}} = \varepsilon$$

$$\Rightarrow \quad T = O(\frac{\beta}{\varepsilon^2})$$

Gradient Descent for Quadratic Functions

$x^* = 0$

Problem: $\min_x \frac{1}{2} x^\top A x$ with $A \in \mathbb{R}^{d \times d}$ being positive-definite.

Theorem: Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and the smallest eigenvalues of A. If we set $\eta \leq \frac{1}{\lambda_{\max}}$, we have

$$\|x_t\|_2 \leq \left(1 - \eta\lambda_{\min}\right)^t \|x_0\|_2$$

$$\begin{aligned}
\|x_{t+1}\|_2 &= \|x_t - \eta A x_t\|_2 \\
&= \|(I - \eta A) x_t\|_2 \\
&\leq \|I - \eta A\|_2 \cdot \|x_t\|_2 \\
&\leq (1 - \eta\lambda_{\min}) \cdot \|x_t\|_2 \\
&\leq (1 - \eta\lambda_{\min})^{t+1} \|x_0\|_2
\end{aligned}$$

If want $\|x_t\| \leq \varepsilon$

$(\because \eta = \frac{1}{\lambda_{\max}}) \Rightarrow$ need $\frac{\lambda_{\max}}{\lambda_{\min}} \log\left(\frac{1}{\varepsilon}\right)$ iters

- $K = \frac{\lambda_{\max}}{\lambda_{\min}}$ condition number
- can be generalized to strongly convex

Momentum: Heavy-Ball Method (Polyak '64)

Problem: $\min_x f(x)$

Method: $v_{t+1} = -\nabla f(x_t) + \beta v_t$

$$x_{t+1} = x_t + \eta v_{t+1}$$

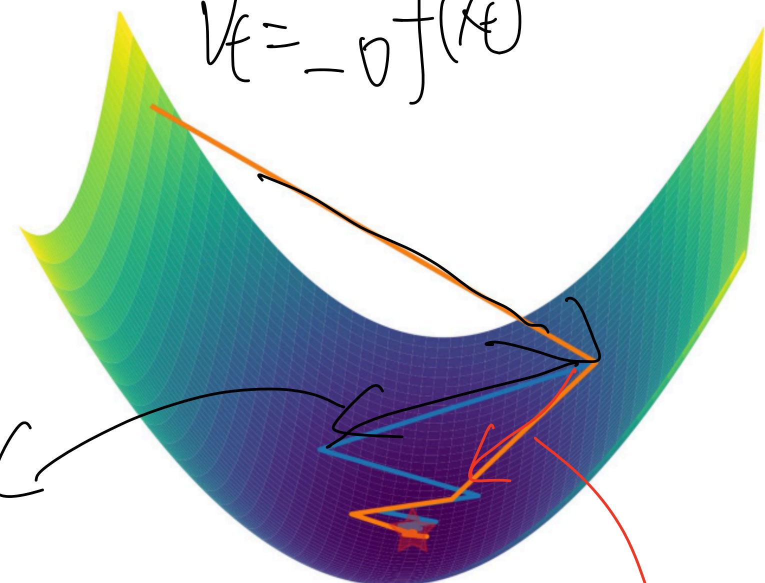
v.s. $K \cdot \log(\frac{1}{\varepsilon})$ using GD

For quadratic optimization,

$\sqrt{K} \cdot \log(\frac{1}{\varepsilon})$

if $K = 10^8 \Rightarrow 10^8 \to 10^4$

$V_t = -\nabla f(x_t)$



$-\nabla f(x_{t+1})$

V_{t+1}

X does not hold for general strongly convex function, with parameter μ

$\forall x,\ x^T \nabla^2 f(x)\, x \geq \mu \cdot \|x\|_2^2$

Momentum: Nesterov Acceleration (Nesterov '89)

Problem: $\min_x f(x)$

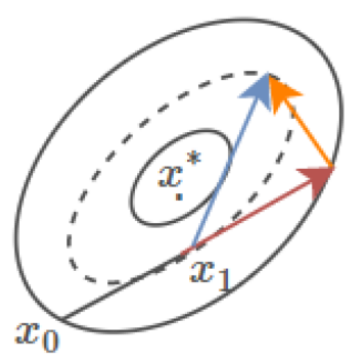
lookahead

Method: $v_{t+1} = -\nabla f(x_t + \beta v_t) + \beta v_t$

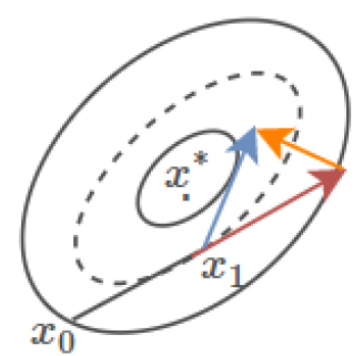
$$x_{t+1} = x_t + \eta v_{t+1}$$

For strongly convex $\Rightarrow$ $\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)$

Polyak's Momentum



Nesterov Momentum



Newton's Method

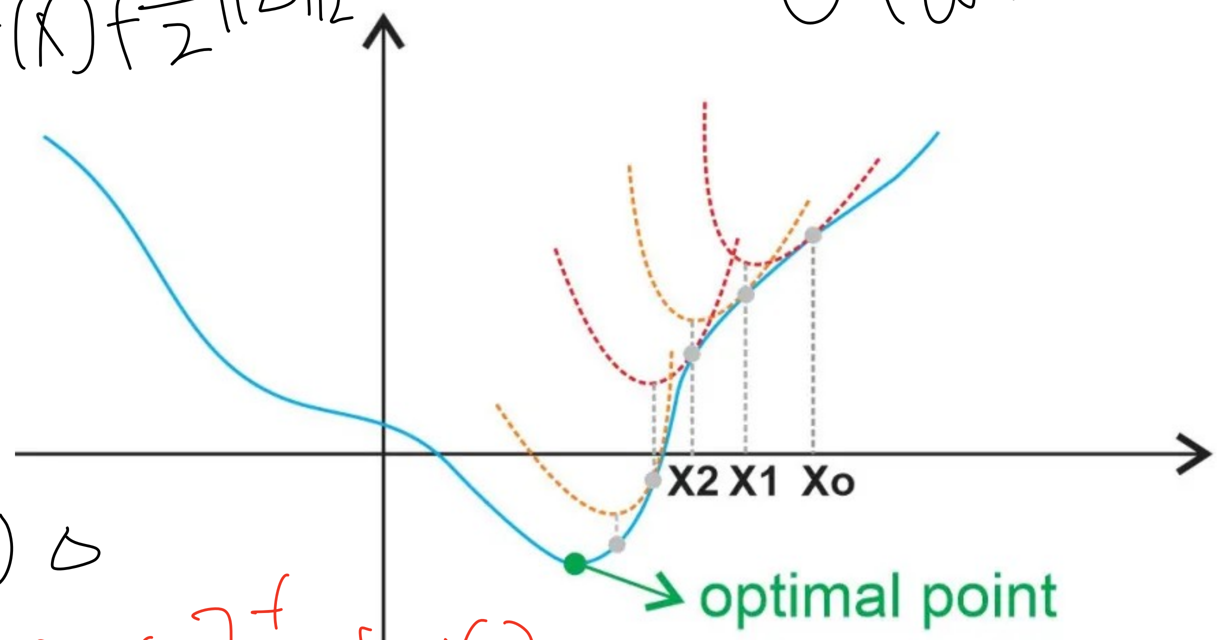
$x_t \in \mathbb{R}^d$

Newton's Method: $x_{t+1} = x_t - \eta(\underbrace{\nabla^2 f(x_t)}_{O(d^2)})^{-1}\underbrace{\nabla f(x_t)}_{O(d)}$

GD: $x_{t+1} = x_t - \eta \cdot \nabla f(x_t)$

$f(x+\Delta) \approx f(x) + \Delta^T \nabla f(x) + \frac{1}{2}\|\Delta\|_2^2$

$\Rightarrow \Delta^* = -\nabla f(x)$



Newton:

$f(x+\Delta) \approx f(x) + \Delta^T \nabla f(x) + \frac{1}{2}\Delta^T \nabla^2 f(x)\, \Delta$

$\Rightarrow \Delta^* = -[\nabla^2 f(x)]^{-1} \nabla f(x)$

Theorem: $O(\log\log(\frac{1}{\varepsilon}))$

AdaGrad (Duchi et al. '11)

diagonal approximation

Newton Method: $x_{t+1} = x_t - \eta(\nabla^2 f(x_t))^{-1} \nabla f(x_t)$

AdaGrad: separate learning rate for every parameter

$$x_{t+1} = x_t - \eta\underbrace{(G_{t+1} + \epsilon I)^{-1}}_{\text{pre-conditioner}} \nabla f(x_t),\ (G_t)_{ii} = \sqrt{\sum_{j=1}^{t-1} \left(\nabla f(x_t)_i\right)^2}$$

- G_t strictly increasing $\Rightarrow$ effective learning rate small
- Comment: default parameter works well $\eta = .01$, $\varepsilon = 10^{-8}$

root - mean - square

RMSProp (Hinton et al. '12)

AdaGrad: separate learning rate for every parameter

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1} \nabla f(x_t), \; (G_t)_{ii} = \sqrt{\sum_{j=1}^{t-1} \left(\nabla f(x_t)_i \right)^2}$$

RMSProp: exponential weighting of gradient norms

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1/2} \nabla f(x_t),$$

$$(G_{t+1})_{ii} = \beta (G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2$$

$$\Leftrightarrow \sum_{t=0}^{T} \beta^{T-t} \left(\nabla f(x_t)_i \right)^2$$

AdaDelta (Zeiler '12)

$\frac{\partial f}{\partial x}$: $\frac{1}{\text{unit of } x}$

$(G_t)^{-1/2}$: unit of x

$\Rightarrow$ Δ : unitless

RMSProp:

$$x_{t+1} = x_t - \eta(G_{t+1} + \epsilon I)^{-1/2} \nabla f(x_t),$$
$$(G_{t+1})_{ii} = \beta(G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2$$

AdaDelta:

$$x_{t+1} = x_t - \eta \Delta x_t,$$
$$\Delta x_t = \sqrt{u_t + \epsilon} \cdot (G_{t+1} + \epsilon I)^{-1/2} \nabla f(x_t)$$
$$(G_{t+1})_{ii} = \rho(G_t)_{ii} + (1-\rho)(\nabla f(x_t)_i)^2,$$
$$u_{t+1} = \rho u_t + (1-\rho)\|\Delta x_t\|_2^2$$

$\sqrt{u_t}$: unit of x

GD : $\Delta \propto \frac{\partial f}{\partial x} \propto \frac{1}{\text{unit of } x}$

Newton : $\Delta \propto \frac{\partial f}{\partial x} \Big/ \frac{\partial^2 f}{\partial x^2} \propto$ unit of x

Adam (Kingma & Ba '14)

AdamW

Momentum:

$$v_{t+1} = -\nabla f(x_t) + \beta v_t, \; x_{t+1} = x_t + \eta v_{t+1}$$

RMSProp: exponential weighting of gradient norms

$$x_{t+1} = x_t - \eta (G_{t+1} + \epsilon I)^{-1} \nabla f(x_t),$$

$$(G_t)_{ii} = \beta (G_t)_{ii} + (1-\beta)(\nabla f(x_t)_i)^2$$

Adam

$$v_{t+1} = \beta_1 v_t + (1-\beta_1) \nabla f(x_t)$$

$$(G_{t+1})_{ii} = \beta_2 (G_t)_{ii} + (1-\beta_2)(\nabla f(x_t)_i)^2$$

$$x_{t+1} = x_t - \eta (G_{t+1} + \epsilon I)^{-1/2} v_{t+1}$$

Default choice nowadays.