

Optimization Methods for Deep Learning



Gradient descent for non-convex optimization

Descent Lemma: Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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:

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

Converging to stationary points

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

Gradient Descent for Quadratic Functions

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

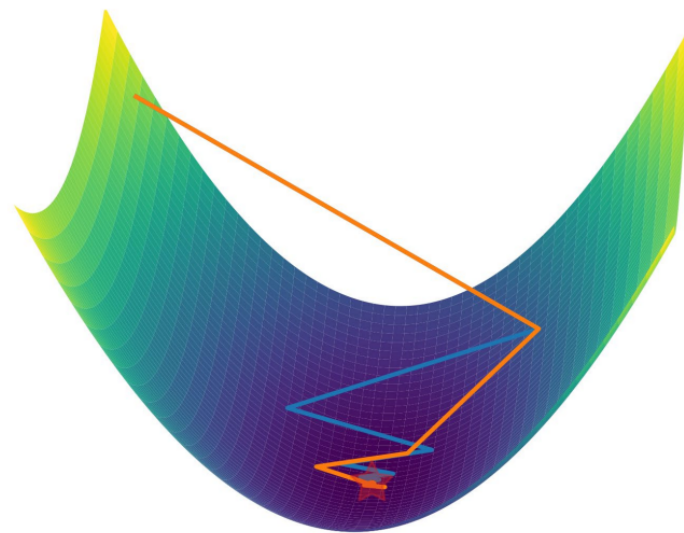
Theorem: Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and the smallest eigenvalues of A . If we set $\eta \leq \frac{1}{\lambda_{\max}}$, we have

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

Momentum: Heavy-Ball Method (Polyak '64)

Problem: $\min_x f(x)$

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

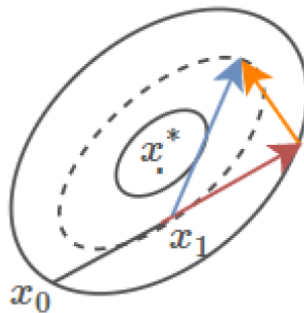


Momentum: Nesterov Acceleration (Nesterov '89)

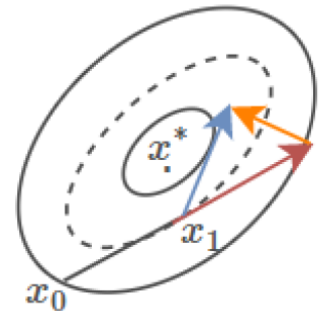
Problem: $\min_x f(x)$

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

Polyak's Momentum

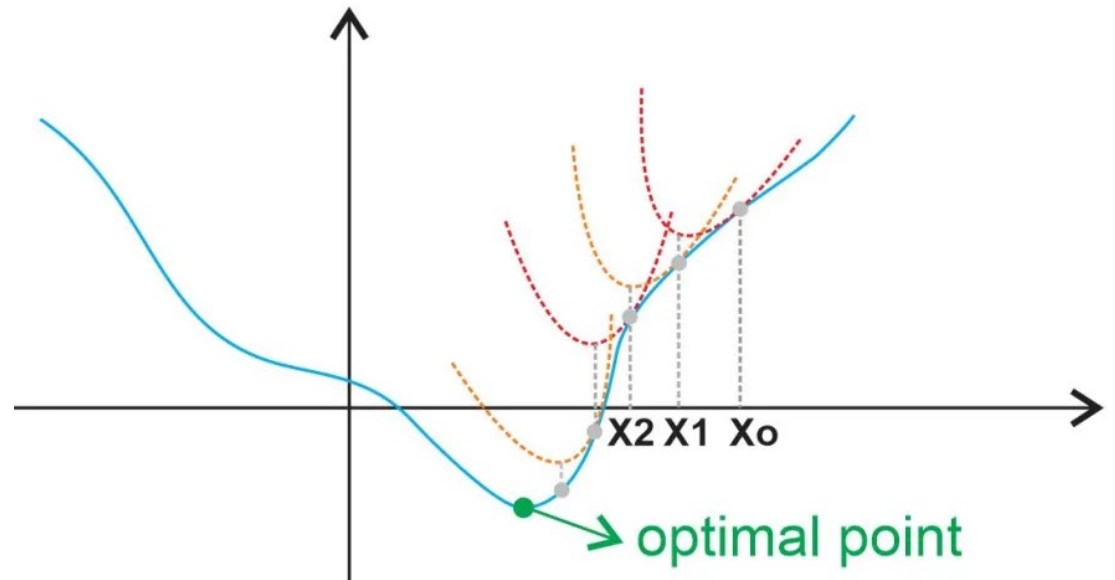


Nesterov Momentum



Newton's Method

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



AdaGrad (Duchi et al. '11)

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

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

RMSProp (Hinton et al. '12)

AdaGrad: separate learning rate for every parameter

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

RMSProp: exponential weighting of gradient norms

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

AdaDelta (Zeiler '12)

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:

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

Adam (Kingma & Ba '14)

Momentum:

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

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.

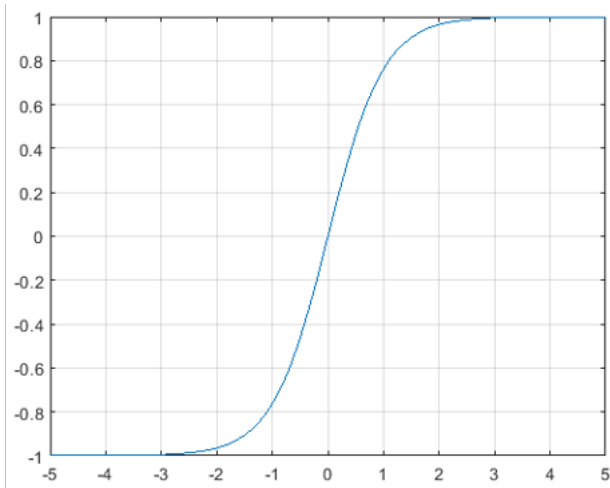
Important Techniques in Neural Network Training



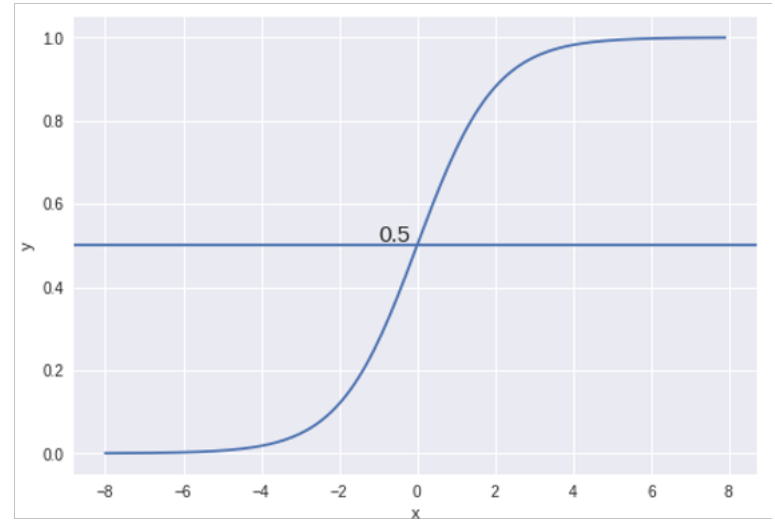
Gradient Explosion / Vanishing

- Deeper networks are harder to train:
 - Intuition: gradients are products over layers
 - Hard to control the learning rate

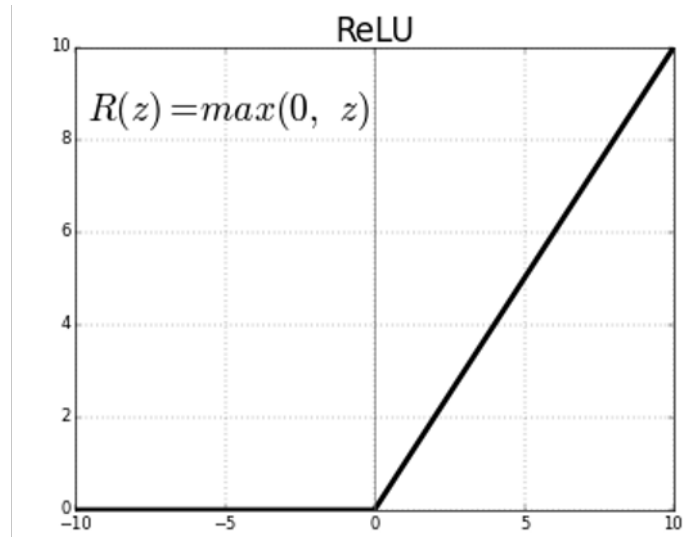
Activation Functions



tanh



sigmoid

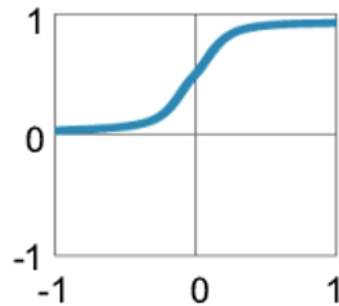


Rectified Linear Unit

Activation Function

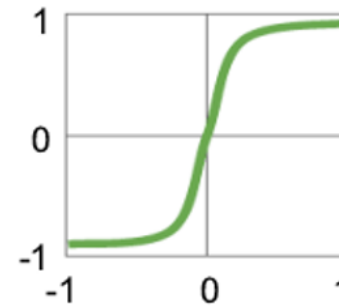
Traditional Non-Linear Activation Functions

Sigmoid



$$y = 1 / (1 + e^{-x})$$

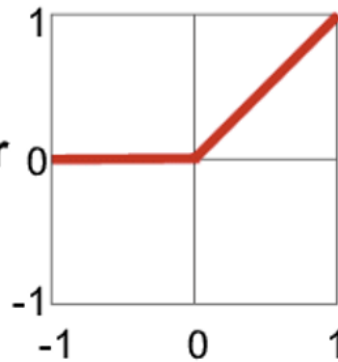
Hyperbolic Tangent



$$y = (e^x - e^{-x}) / (e^x + e^{-x})$$

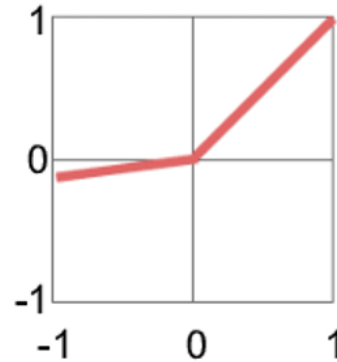
Modern Non-Linear Activation Functions

Rectified Linear Unit (ReLU)



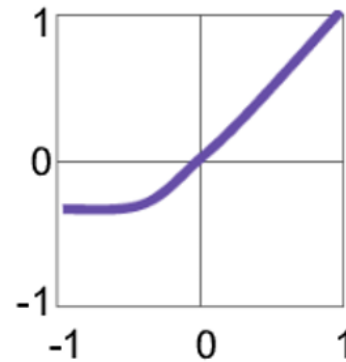
$$y = \max(0, x)$$

Leaky ReLU



$$y = \max(\alpha x, x)$$

Exponential LU



$$y = \begin{cases} x, & x \geq 0 \\ \alpha(e^x - 1), & x < 0 \end{cases}$$

α = small const. (e.g. 0.1)

Initialization

- Zero-initialization
- Large initialization
- Small initialization

- Design principles:
 - Zero activation mean
 - Activation variance remains same across layers

Kaiming Initialization (He et al. '15)

- $W_{ij}^{(h)} \sim \mathcal{N}\left(0, \frac{2}{d_h}\right)$.
- $b^{(h)} = 0$
- Designed for ReLU activation
- 30-layer neural network

