Optimizations on Gradient Descent

Dr. Chandranath Adak

Dept. of CSE, Indian Institute of Technology Patna

October 8, 2025

Gradient Descent (GD) with Momentum

A few observations on GD

- GD takes significant time to navigate regions having a gentle slope due to
  - The gradient in these regions is very small
  - Learning rate does not help
- Can we do something better?

Gradient Descent (GD) with Momentum

A few observations on GD

- GD takes significant time to navigate regions having a gentle slope due to
  - The gradient in these regions is very small
  - Learning rate does not help
- Can we do something better?

Gradient Descent with Momentum: Intuition

In addition to the current update, also consider the update-history

Gradient Descent (GD) with Momentum

A few observations on GD

- GD takes significant time to navigate regions having a gentle slope due to
  - The gradient in these regions is very small
  - Learning rate does not help
- Can we do something better?

Gradient Descent with Momentum: Intuition

In addition to the current update, also consider the update-history

Gradient Descent with Momentum

$$\mu_t = \beta \mu_{t-1} + \alpha \nabla w_t$$

$$w_{t+1} = w_t - \mu_t$$

Gradient Descent (GD) with Momentum

Gradient Descent with Momentum

$$\mu_t = \beta\mu_{t-1} + \alpha\nabla w_t$$

$$w_{t+1} = w_t - \mu_t$$

$$\begin{aligned}
\mu_0 =& 0 \\
\mu_1 =& \beta\mu_0 + \alpha\nabla w_1 = & \alpha\nabla w_1 \\
\mu_2 =& \beta\mu_1 + \alpha\nabla w_2 = & \beta\alpha\nabla w_1 + \alpha\nabla w_2 \\
\mu_3 =& \beta\mu_2 + \alpha\nabla w_3 = & \beta^2\alpha\nabla w_1 + \beta\alpha\nabla w_2 + \alpha\nabla w_3 \\
\vdots \\
\mu_t =& \beta\mu_{t-1} + \alpha\nabla w_t = & \underbrace{\beta^{t-1}\alpha\nabla w_1 + \beta^{t-2}\alpha\nabla w_2 + \ldots + \beta\alpha\nabla w_{t-1}}_{\text{More weight on recent history, less weight on old history}} + \alpha\nabla w_t
\end{aligned}$$

Gradient Descent (GD) with Momentum

Hyper-parameter for Momentum (A heuristic)

The following schedule was suggested by Sutskever et al., 2013

$$\beta_t = \min(1 - 2^{-1-\log_2(\lfloor t/250 \rfloor + 1)}, \beta_{max})$$

where, β_{max} was chosen from $\{0.999, 0.99, 0.9, 0\}$

$$\beta_0 = 0.5$$
$$\beta_{250} = 0.75$$
$$\beta_{750} = 0.875$$
$$\beta_{1750} = 0.9375$$

Observation

As the step increases, β_t also increases up to β_{max}

What next?

Limitations of Gradient Descent (GD) with Momentum

- GD with momentum can take large steps in the regions having gentle slopes
- Is moving fast always good?
  - It oscillates in and out around the region of minima as the momentum carries it out of the region

Nesterov Accelerated Gradient Descent: Intuition

- In GD with momentum, two factors responsible for updation

$$w_{t+1} = w_t - \underbrace{\beta\mu_{t-1}}_{\text{update-history}} + \underbrace{\alpha\nabla w_t}_{\text{current update}}$$

What next?

Limitations of Gradient Descent (GD) with Momentum

- GD with momentum can take large steps in the regions having gentle slopes
- Is moving fast always good?
  - It oscillates in and out around the region of minima as the momentum carries it out of the region

Nesterov Accelerated Gradient Descent: Intuition

- In GD with momentum, two factors responsible for updation

$$w_{t+1} = \underbrace{\left(w_t - \underbrace{\beta \mu_{t-1}}_{\text{update-history}} \right)}_{\color{red}{\text{Why not check for update at this point?}}} + \underbrace{\alpha \nabla w_t}_{\text{current update}}$$

Nesterov Accelerated Gradient Descent

Nesterov Accelerated Gradient Descent: Intuition

- In GD with momentum, two factors responsible for updation

$$w_{t+1} = \underbrace{\left(w_t - \underbrace{\beta \mu_{t-1}}_{\text{update-history}} \right)}_{\text{Look-ahead (LA) and check for update}} + \underbrace{\alpha \nabla w_t}_{\text{current update}}$$

Nesterov Accelerated Gradient Descent

$$w_{LA}^{t} = w_t - \beta\ \mu_{t-1}$$

$$\mu_t = \beta\ \mu_{t-1} + \alpha\ \nabla w_{LA}^{t}$$

$$w_{t+1} = w_t - \mu_t$$

Adaptive Learning Rate

Step Decay

- Learning rate (α_t) is a function of no. of steps (t)
- Start with a comparatively large initial learning rate, decay the learning rate after a specific step-interval

- Two parameters need to be decided
- Step-interval
  - Step-interval can be a fixed value
  - Step-interval can depend on validation error
    - Decay the learning rate after an epoch if the validation error is more than the one at the end of the previous epoch
- Decay rate
  - After each step-interval, the learning rate can be half of itself
  - $\alpha_t = \frac{\alpha_0}{1+kt}$; k is another hyper-parameter

Adaptive Learning Rate

Exponential Decay

$$\alpha_t = \alpha_0^{-kt}$$

k is a hyper-parameter; t is the step number

- These are all heuristic strategies
- There is no best strategy

Adagrad

- Decay the learning rate for parameters in proportion to their update history

Adagrad

$$v_t = v_{t-1} + \left(\nabla w_t\right)^2$$

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \nabla w_t$$

RMSProp

- Adagrad decays the learning rate very aggressively
- After a few updates, the frequent parameters start receiving very smaller updates
- **Motivation for RMSProp**: Control the rapid decay of learning rate for Adagrad
- In practice, $\beta = 0.999$

RMSProp

$$v_t = \beta \; v_{t-1} + (1 - \beta) \; (\nabla w_t)^2$$

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \; \nabla w_t$$

ADAM

- Combination of RMSProp and GD with momentum
- In practice, $\beta_1 = 0.999$ and $\beta_2 = 0.9$

ADAM

$$v_t = \beta_1 \; v_{t-1} + \left(1 - \beta_1\right) \; \left(\nabla w_t\right)^2$$

$$\mu_t = \beta_2 \; \mu_{t-1} + \left(1 - \beta_2\right) \; \nabla w_t$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_1^t}; \quad \hat{\mu}_t = \frac{\mu_t}{1 - \beta_2^t}$$

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t + \epsilon}} \; \hat{\mu}_t$$

Bias Correction

- μ_t is the exponentially moving average of the gradient
- The motivation of using momentum was
  - Instead of relying only on the current gradient, can we consider the overall behaviour of the gradients over earlier timesteps?
- Essentially, we are interested in the expected value of the gradients
- Ideally, $E(\nabla w_t) = E(\mu_t)$

Bias Correction

$$
\begin{aligned}
\mu_t &= \beta_2 \ \mu_{t-1} + (1-\beta_2) \ \nabla w_t \\
\mu_0 &= 0 \\
\mu_1 &= \beta_2 \ \mu_0 + (1-\beta_2) \ \nabla w_1 \\
&= (1-\beta_2) \ \nabla w_1 \\
\mu_2 &= \beta_2 \ \mu_1 + (1-\beta_2) \ \nabla w_2 \\
&= \beta_2 \ (1-\beta_2) \ \nabla w_1 + (1-\beta_2) \ \nabla w_2 \\
\mu_3 &= \beta_2 \ \mu_2 + (1-\beta_2) \ \nabla w_3 = \\
&= \beta_2^2 \ (1-\beta_2) \ \nabla w_1 + \beta_2 \ (1-\beta_2) \ \nabla w_2 + (1-\beta_2) \ \nabla w_3 \\
&\vdots \\
\mu_t &= \beta_2 \ \mu_{t-1} + (1-\beta_2) \ \nabla w_t = \\
&= \beta_2^{t-1} \ (1-\beta_2) \ \nabla w_1 + \beta_2^{t-2} \ (1-\beta_2) \ \nabla w_2 + \ldots + (1-\beta_2) \ \nabla w_t \\
&= \sum_{i=1}^{t} \beta_2^{t-i} \ (1-\beta_2) \ \nabla w_i = (1-\beta_2) \ \sum_{i=1}^{t} \beta_2^{t-i} \ \nabla w_i
\end{aligned}
$$

Bias Correction

$$\mu_t = (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \nabla w_i$$

$$E\left[\mu_t\right] = E\left[(1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \nabla w_i\right]$$

$$E\left[\mu_t\right] = (1-\beta_2)\ E\left[\sum_{i=1}^{t} \beta_2^{t-i} \nabla w_i\right]$$

$$E\left[\mu_t\right] = (1-\beta_2) \sum_{i=1}^{t} E\left[\beta_2^{t-i} \nabla w_i\right]$$

$$E\left[\mu_t\right] = (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i}\ E\left[\nabla w_i\right]$$

Assumption

All ∇w_i follows the same distribution, i.e., $E[\nabla w_i] = E[\nabla w]$

Bias Correction

$$E[\mu_t] = (1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}\,E[\nabla w_i]$$

$$E[\mu_t] = (1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}\,E[\nabla w]$$

$$E[\mu_t] = E[\nabla w](1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}$$

$$E[\mu_t] = E[\nabla w](1-\beta_2)\left(\beta_2^{t-1}+\beta_2^{t-2}+\ldots+\beta_2^{1}+\beta_2^{0}\right)$$

$$E[\mu_t] = E[\nabla w](1-\beta_2)\,\frac{1-\beta_2^t}{1-\beta_2}$$

$$E[\nabla w] = \frac{E[\mu_t]}{1-\beta_2^t}$$

$$E[\nabla w] = E\left[\frac{\mu_t}{1-\beta_2^t}\right]$$

$$E[\nabla w] = E[\hat{\mu}_t] \qquad \text{therefore, } \hat{\mu}_t = \frac{\mu_t}{1-\beta_2^t}$$

Thank You!