**2.2.3. Một số thuật toán tối ưu khác**

***Nesterov accelerated gradient***

Nesterov accelerated gradient (Nesterov, 1983, 2004) là một biến thể nhỏ của kỹ thuật momentum cho phép gradients thoát các điểm local minimum, khiến thuật toán hội tụ nhanh hơn.

Ý tưởng cơ bản là dự đoán hướng đi trong tương lai, tức nhìn trước một bước. Cách thức này có 1 ít khác biệt so với momentum update, với momentum update ta tính toán đạo hàm tại vị trí hiện hành rồi sau đó làm 1 cú nhảy tới dựa trên vector momentum trước đó, với Nesterov momentum thay vì tính toán đạo hàm tại điểm hiện hành, chúng ta dựa vào vector momentum cũ để tính toán vị trí sắp tới, rồi sau đó mới dùng đạo hàm tại vị trí mới để correct lại. Theo dõi hình dưới đây:



Ý tưởng của Nesterov accelerated gradient. (Nguồn: [CS231n Stanford: Convolutional Neural Networks for Visual Recognition](https://cs231n.github.io/neural-networks-3/))

* Với momentum thông thường: lượng thay đổi là tổng của hai vector: momentum vector và gradient ở thời điểm hiện tại.
* Với Nesterove momentum: lượng thay đổi là tổng của hai vector: momentum vector và gradient ở thời điểm được xấp xỉ là điểm tiếp theo.

Công thức cập nhật của NAG được cho như sau:

***Adagrad***

Adagrad (Duchi, 2011) cho phép tinh chỉnh các gradients khác nhau bằng  
việc có learning rate riêng biệt *η*. Cho mỗi giá trị của *f, f’, i*

Không giống như các cách thức trước, learning rate hầu như giống nhau cho quá trình learning, adagrad coi learning rate cũng là một tham số. Nó tạo các update lớn với các dữ liệu khác biệt nhiều và các update nhỏ cho các dữ liệu ít khác biệt.

Adagrad chia learning rate với tổng bình phương của lịch sử biến thiên (đạo hàm). Quy luật cập nhật như sau:

Trong đó:

ϵ là hệ số để tránh lỗi chia cho 0, default là 1e−8

là một ma trận chéo nơi mà mỗi phần tử (i, i) là bình phương của đạo hàm vector tham số tại thời điểm t

Một lợi ích dễ thấy của Adagrad là tránh việc điều chỉnh learning rate bằng tay, thường sẽ để default là 0.01 và thuật toán sau đó sẽ tự động điều chỉnh.

Còn điểm yếu của Adagrad là tổng bình phương biến thiên sẽ lớn dần theo thời gian cho đến khi nó làm learning rate cực kỳ nhỏ, làm việc training trở nên đóng băng.

***RMSprop***

RMSprop (Hinton, 2012) là một kỹ thuật điều chỉnh learning rate có tính thích nghi, nó chưa từng được công bố, và được đề xuất bởi Geoff Hinton trong bài giảng trên Coursera.

Vì Adagrad thêm vào gradient từ epoch đầu tiên, các trọng số bị buộc phải giảm  
đơn điệu. Điều này có thể được giải quyết thông qua kỹ thuật Adadelta:  
với *γ* là một tham số mới, thường là 0*.*9. Công thức Adadelta theo Adagrad như  
sau  
*η* được thiết lập một lần và cho tất cả (thường là 10*−*3).

***Adadelta***

Adadelta (Zeiler, 2012) là một phần mở rộng của RMSprop, Adadelta sinh ra để làm giảm nhược điểm của Adagrad (việc làm thay đổi learning rate theo tính đơn điệu giảm). Nó giới hạn sự tích lũy của độ biến thiên tới một giới hạn nhất định.

Để làm được điều trên nó đưa ra khái niệm *running average* phụ thuộc vào trung bình trước và độ dốc hiện tại.

và quy luật cập nhật mới như sau:

***Adam***

Adam (Kingma & Ba, 2014) thường là sự lựa chọn mặc định đối với một số thuật toán deep learning. Giống với Adadelta và RMSprop, nó duy trì trung bình bình phương độ dốc (slope) quá khứ vt và cũng đồng thời duy trì trung bình độ dốc quá khứ mt, giống momentum.

, ,

với *β*1 và *β*2 thường là 0*.*9 và 0*.*999. Nhưng sức mạnh và tính chắc chắn của  
Adam nằm ở việc nó làm cho toàn bộ quá trình học tập phụ thuộc vào yếu tố  
precise. Để tránh những vấn đề về số trong các bước đầu tiên, các vector này  
được thay đổi tỷ lệ:

,

trước khi vào các quy tắc cập nhật:

Đây là kỹ thuật tối ưu hoá được sử dụng trong bài báo cáo này, cùng với độ giảm  
learning rate:

*α*0 được xác định bởi cross-validation, và *η*0 thường bắt đầu trong khoảng 10*−*3 *−*  
10*−*2.

Trong khi momentum giống như một quả cầu lao xuống dốc, thì Adam lại giống như một quả cầu rất nặng và có ma sát (friction), nhờ vậy nó dễ dàng vượt qua local minimum và đạt tới điểm tối ưu nhất (flat minimum).

<https://viblo.asia/p/thuat-toan-toi-uu-adam-aWj53k8Q56m>