

Huấn luyện mạng nơ-ron nhiều tầng ẩn bằng thuật toán Adam

Nhóm sinh viên thực hiện:

- Nguyễn Ngọc Lan Như - 1712644
- Hoàng Minh Quân – 1712688

Giáo viên hướng dẫn: Th.S. Trần Trung Kiên

Mục lục

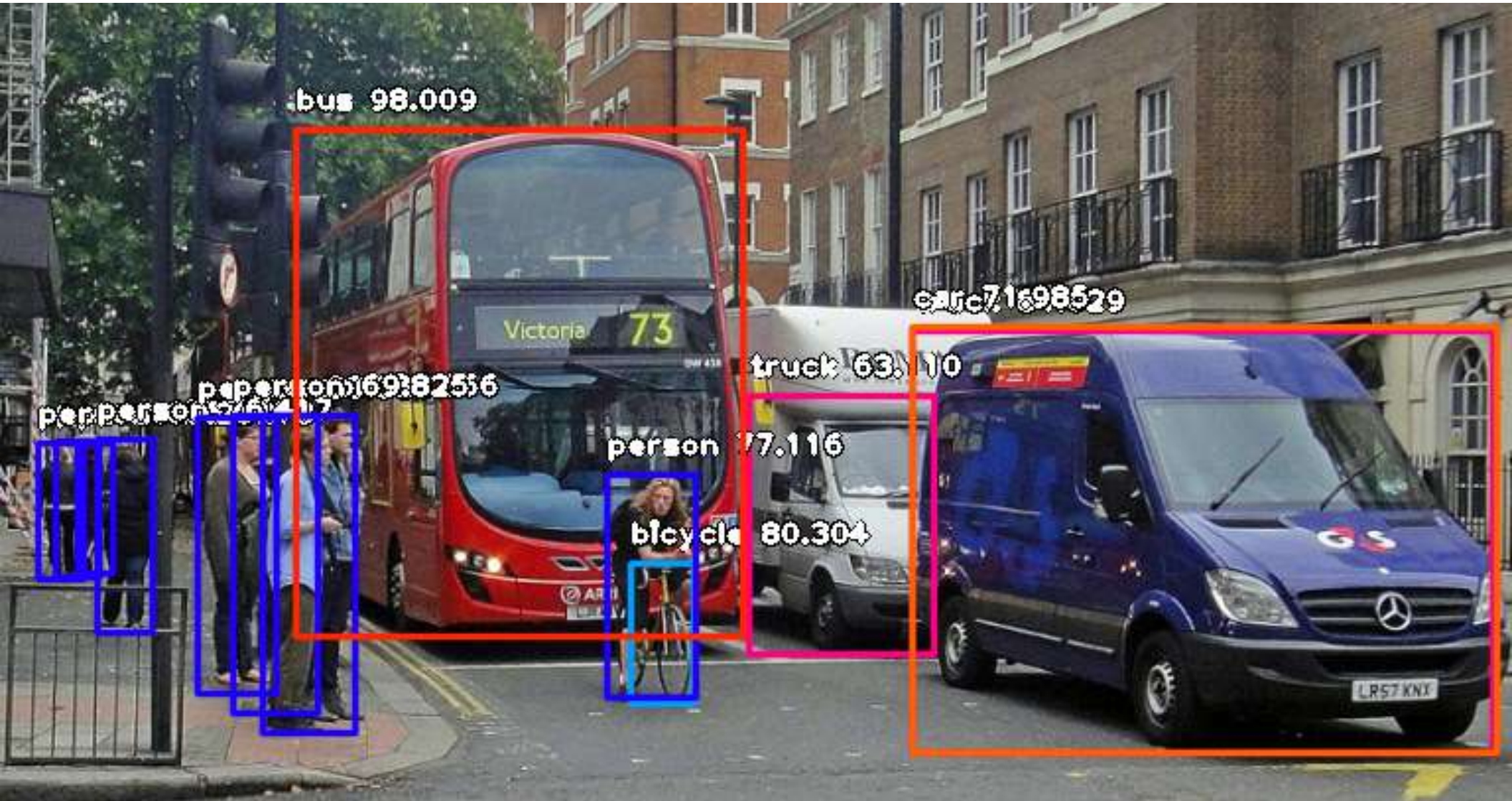
1. Giới thiệu đề tài
2. Kiến thức nền tảng
3. Thuật toán Adam
4. Thí nghiệm
5. Tổng kết

1.

Giới thiệu đề tài

Giới thiệu đề tài

Tại sao lại sử dụng mạng nơ-ron nhiều tầng ẩn?



Giới thiệu đề tài

Tại sao lại sử dụng mạng nơ-ron nhiều tầng ẩn?

MACHINE TRANSLATION



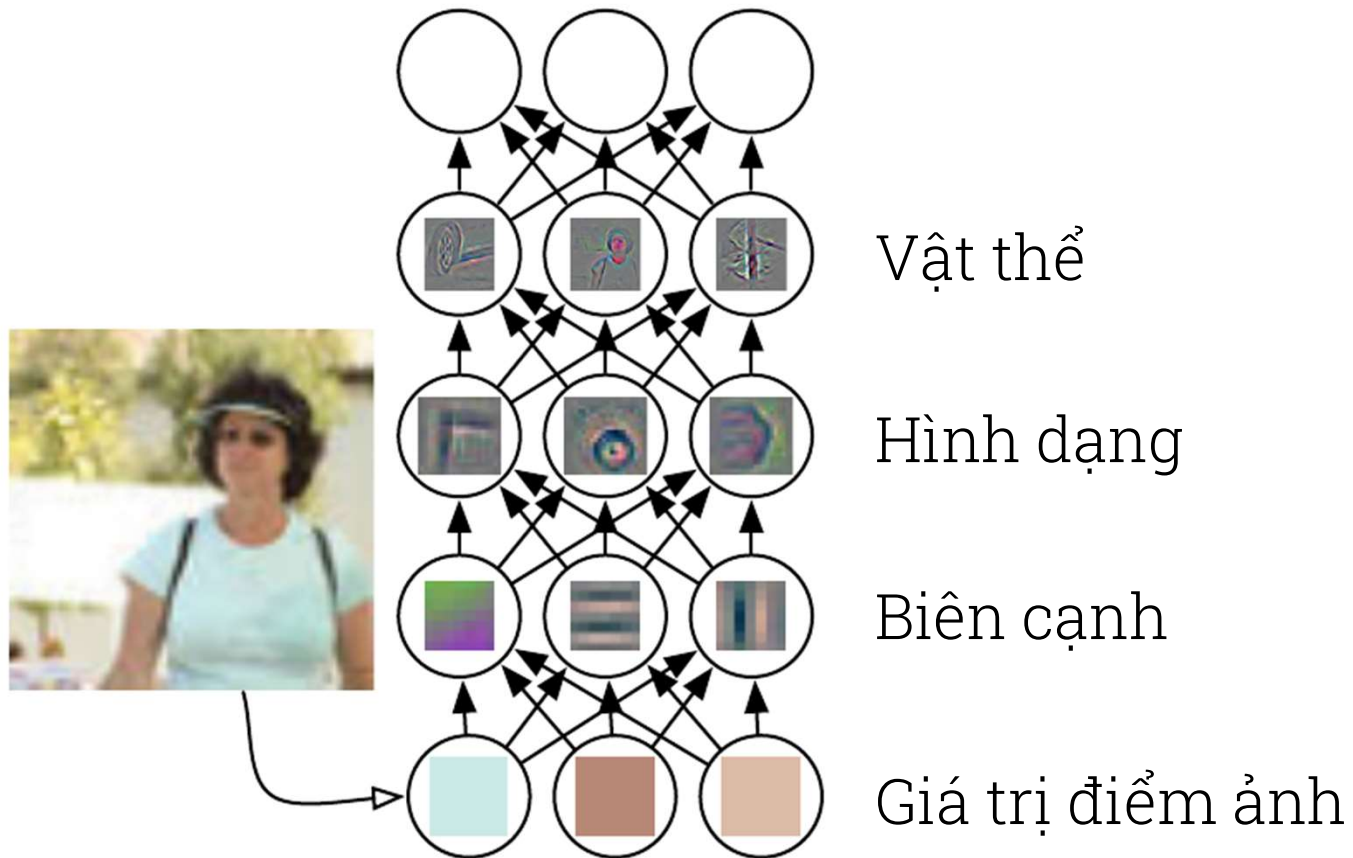
Giới thiệu đề tài

Tại sao lại sử dụng mạng nơ-ron nhiều tầng ẩn?



Giới thiệu đề tài

Tại sao lại sử dụng mạng nơ-ron nhiều tầng ẩn?



Giới thiệu đề tài

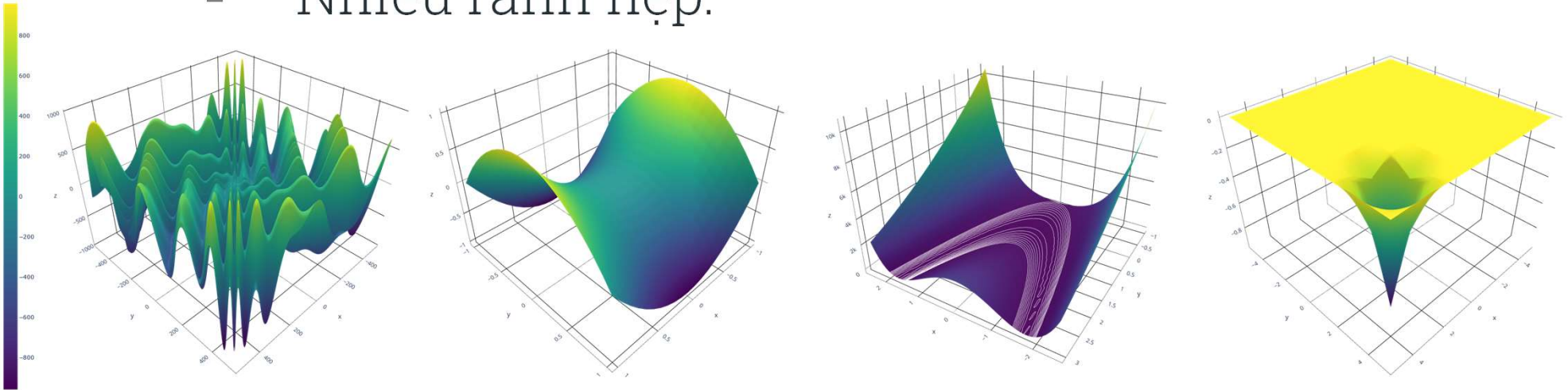
Bài toán huấn luyện mạng nơ-ron nhiều tầng ẩn

- **Input:** Hàm chi phí với tham số là các trọng số của mạng nơ-ron nhiều tầng ẩn. Hàm chi phí cho biết sự sai lệch giữa kết quả dự đoán của mạng nơ-ron so với giá trị đúng trên tập dữ liệu huấn luyện, hay *độ lỗi*.
- **Output:** Bộ trọng số của mạng nơ-ron nhiều tầng ẩn cho độ lỗi là nhỏ nhất, hoặc đủ nhỏ.

Giới thiệu đề tài

Thách thức

- Mặt phẳng lỗi phức tạp:
 - Nhiều cực tiểu địa phương.
 - Nhiều điểm yên ngựa.
 - Nhiều vùng bằng phẳng.
 - Nhiều rãnh hẹp.



*Yann Dauphin et al., "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization", *Advances in Neural Information Processing Systems* 27, 2014, pp. 2933-2941.

Giới thiệu đề tài

Đề tài liên quan: GD

- Hướng tiếp cận truyền thống: Gradient Descent (GD)
 - Sử dụng gradient của cả tập dữ liệu để xác định hướng đi có sự thay đổi lớn nhất.
 - Chi phí tính toán cao

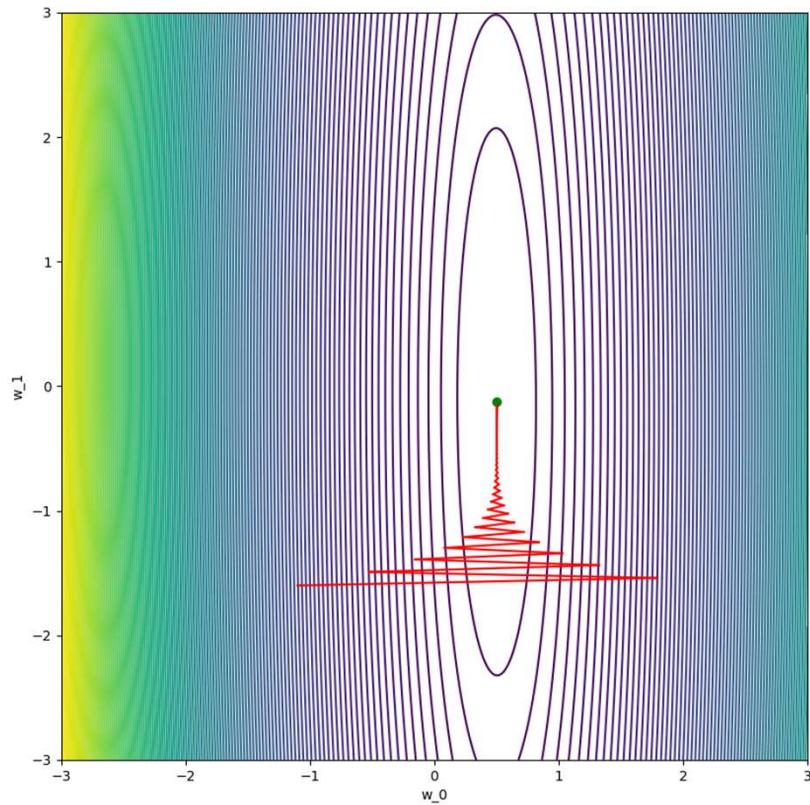
Giới thiệu đề tài

Đề tài liên quan: SGD

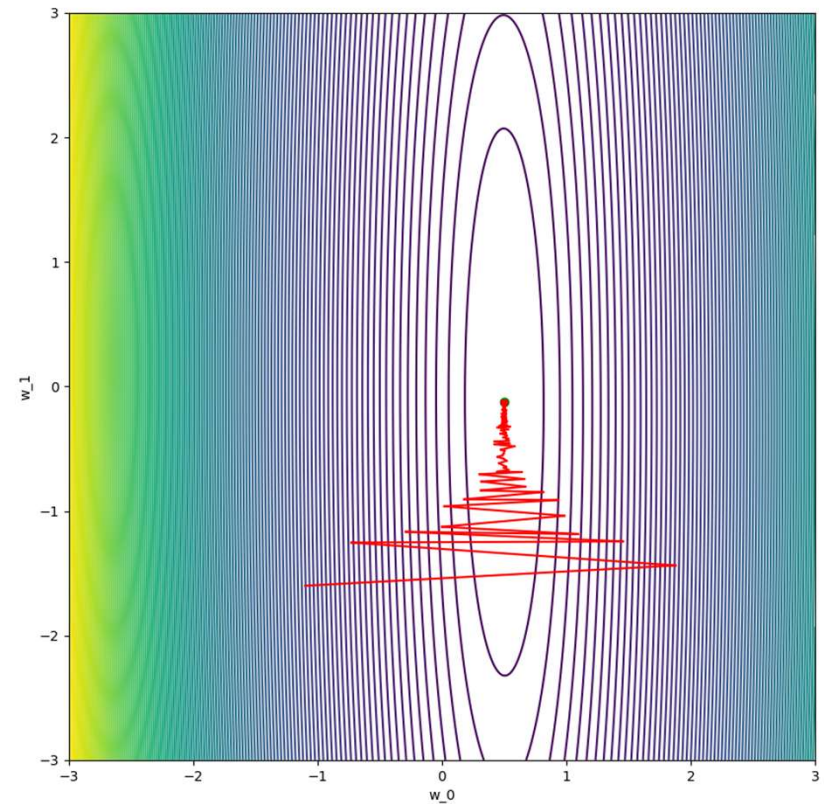
- Hướng tiếp cận truyền thống: **Stochastic Gradient Descent** (SGD)
 - Sử dụng gradient của **một tập con** của dữ liệu để **xấp xỉ** hướng của gradient trên cả tập dữ liệu.

Giới thiệu đề tài

Đề tài liên quan: SGD



Gradient Descent



Stochastic Gradient Descent

Giới thiệu đề tài

Đề tài liên quan: Momentum

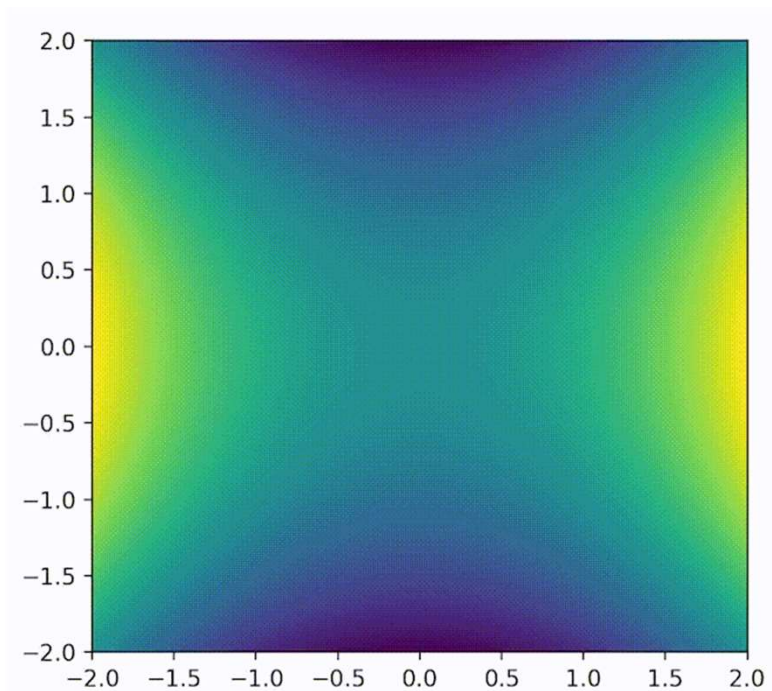
- Cải tiến: Stochastic Gradient Descent **with Momentum** (Momentum)*.
 - Tăng độ lớn cập nhật khi chiều gradient không đổi.
 - Giảm độ lớn cập nhật khi chiều gradient thay đổi liên tục.

*Ning Qian, "On the momentum term in gradient descent learning algorithms", *Journal of the International Neural Network Society*, 1999, vol. 12, pp. 145-151.

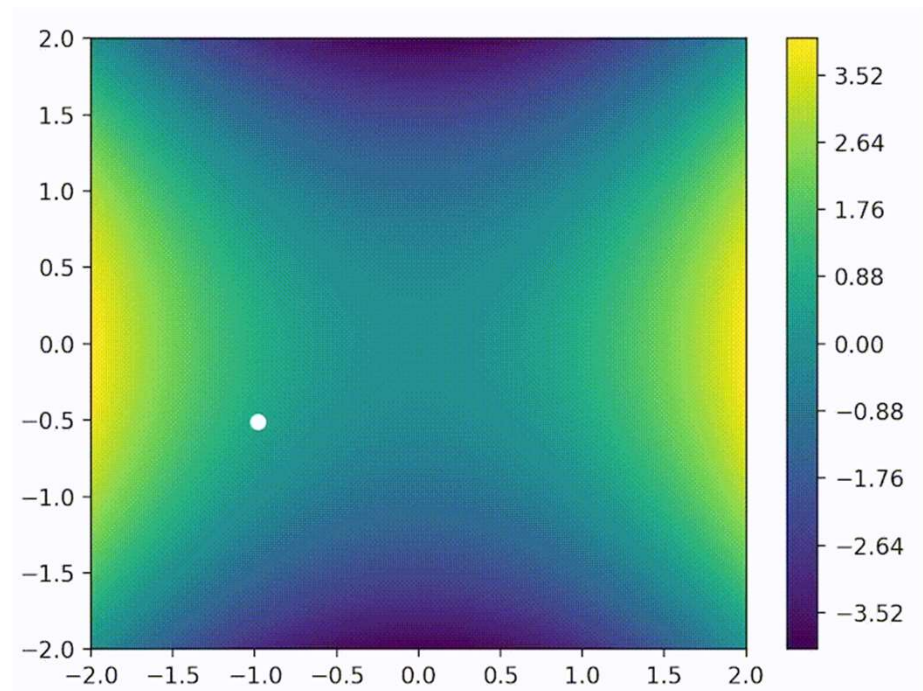
Giới thiệu đề tài

Đề tài liên quan: Momentum

- Di chuyển nhanh hơn tại các vùng bằng phẳng



Momentum

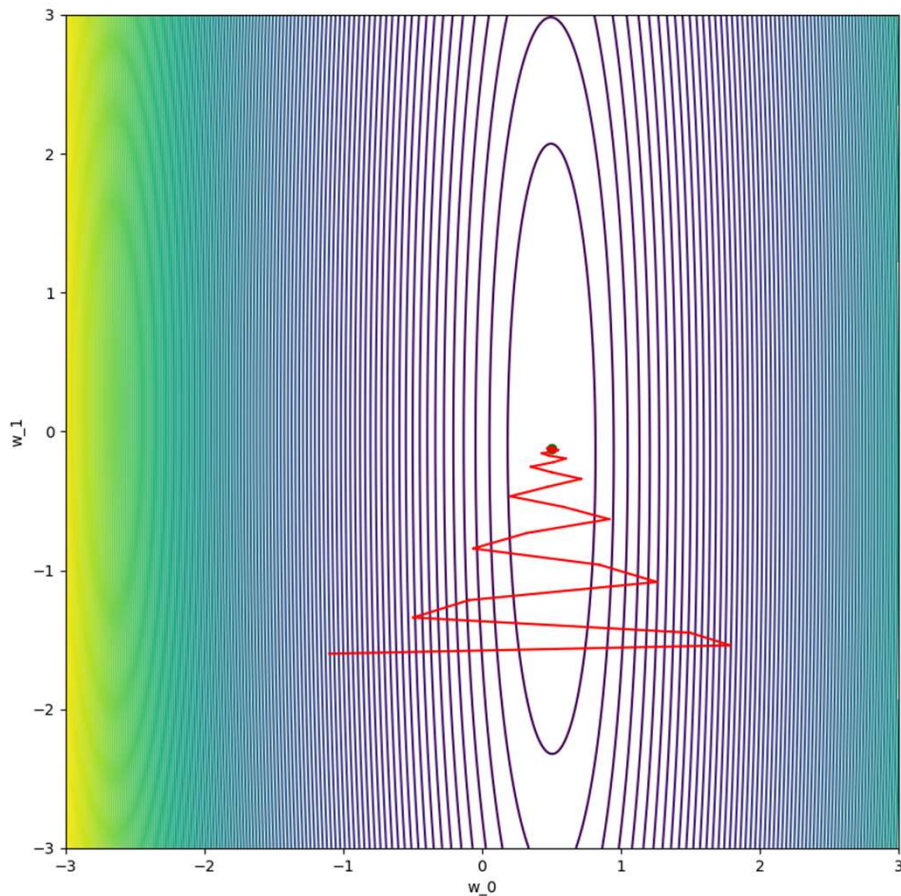


GD

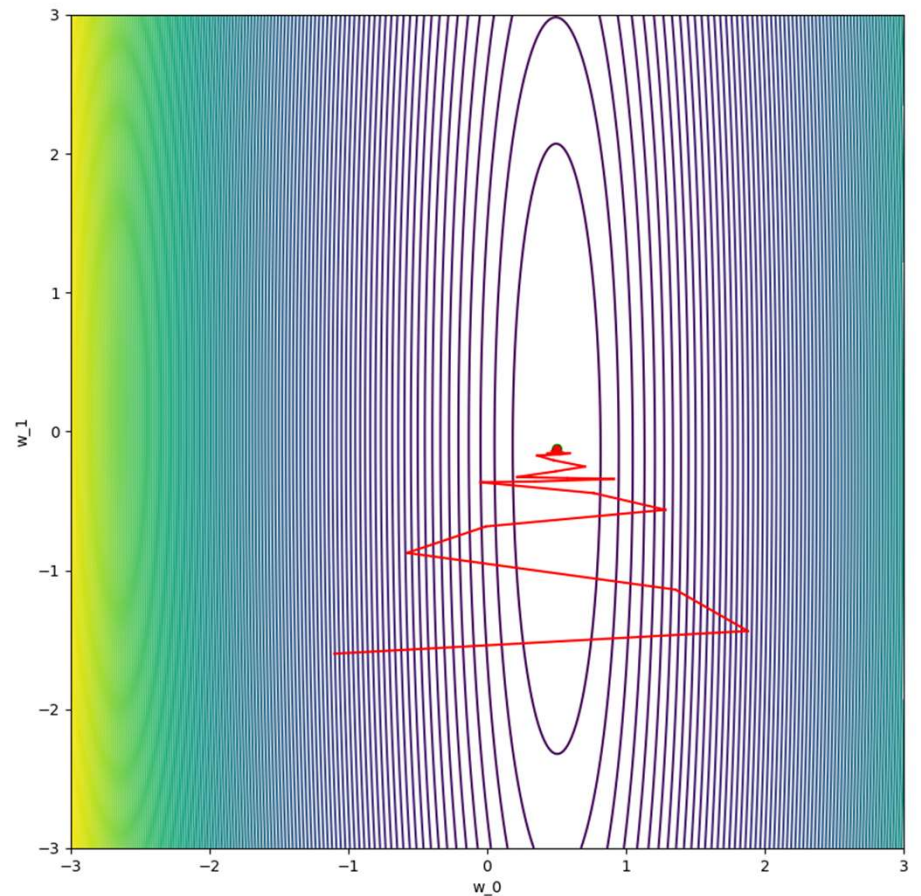
Giới thiệu đề tài

Đề tài liên quan: Momentum

- Xấp xỉ gradient của cả tập dữ liệu tốt hơn.



GD with Momentum

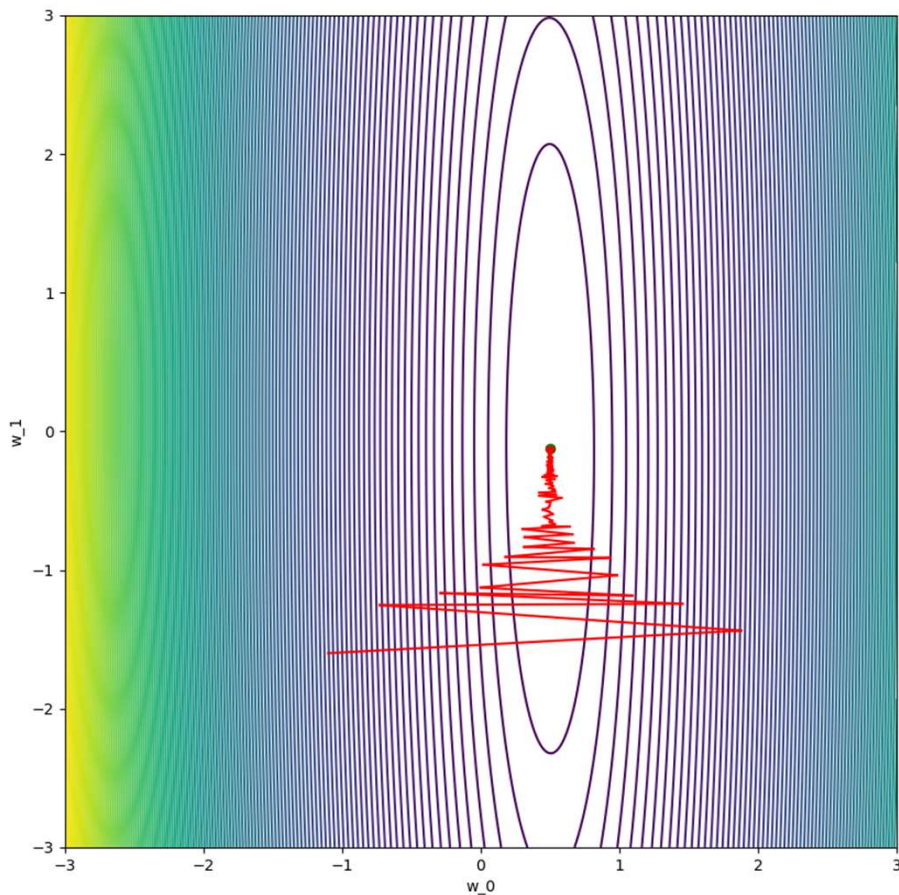


SGD with Momentum

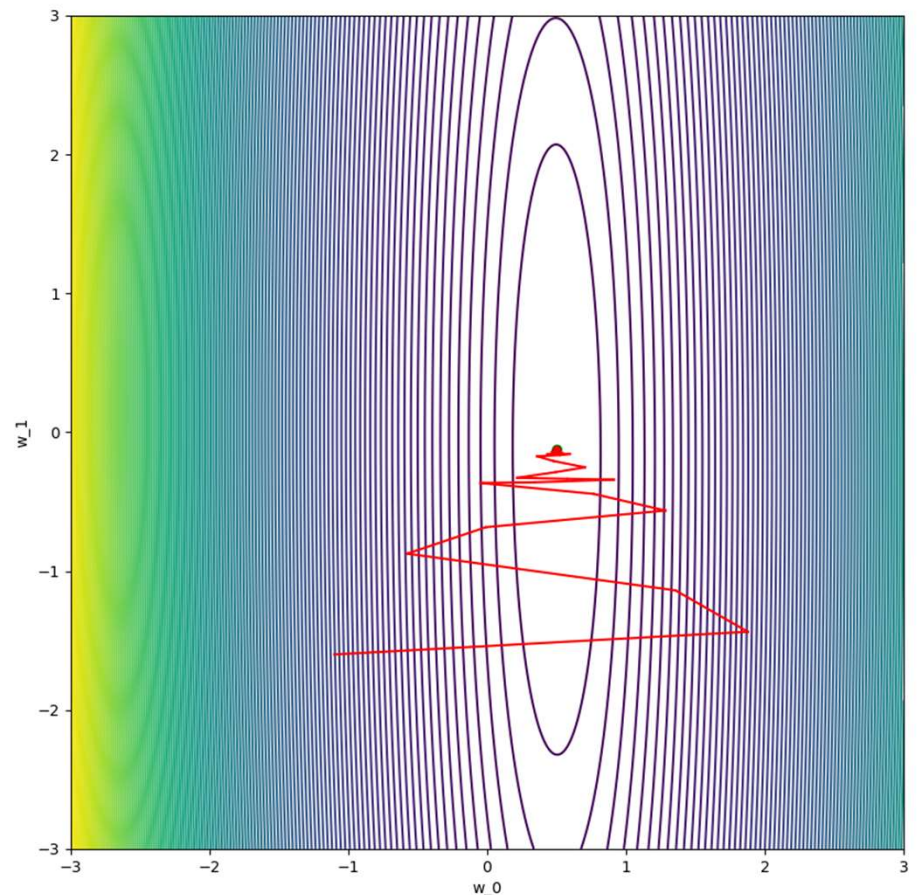
Giới thiệu đề tài

Đề tài liên quan: Momentum

- Giảm dao động.



SGD

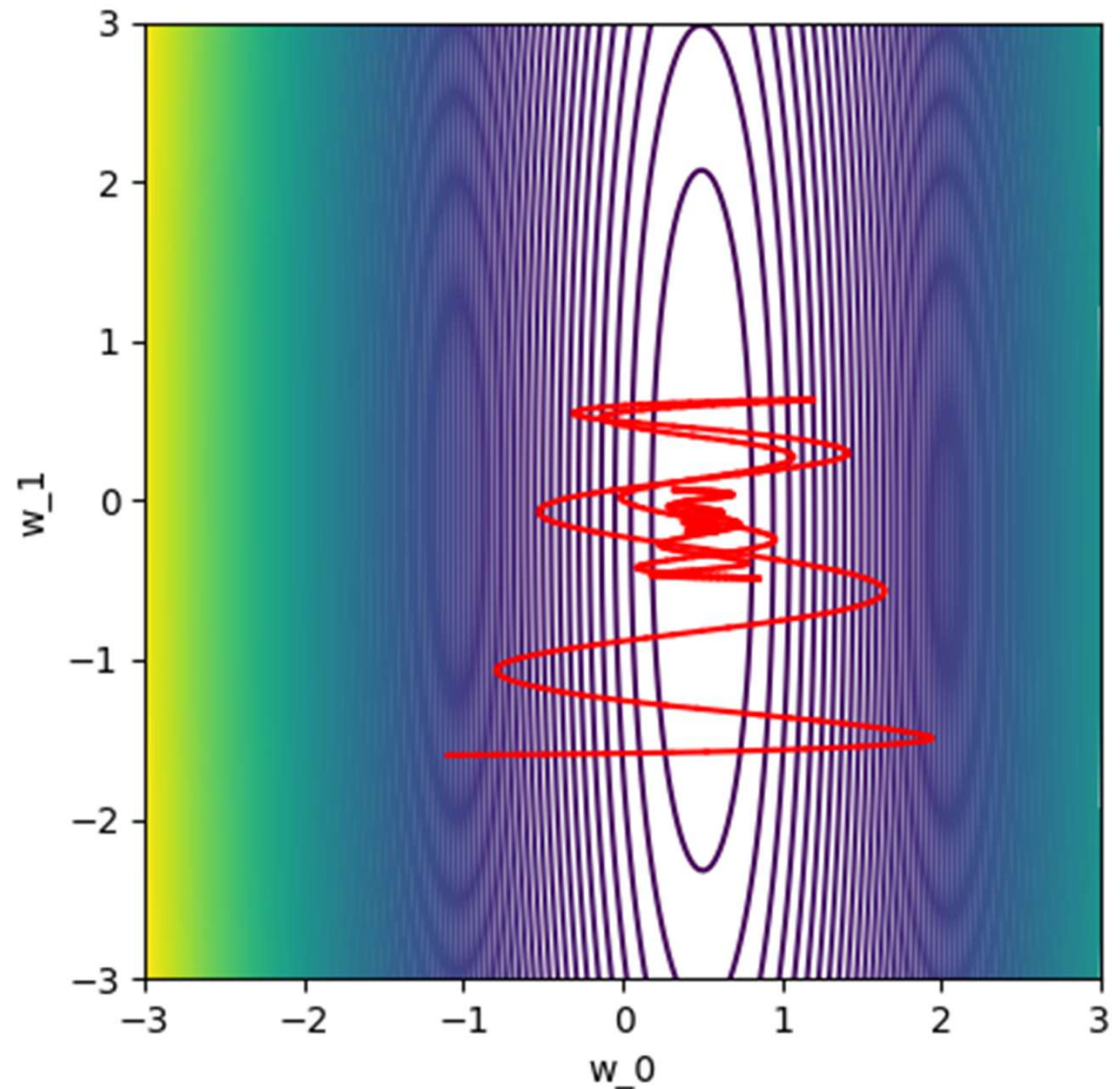


SGD with Momentum

Giới thiệu đề tài

Đề tài liên quan: Momentum

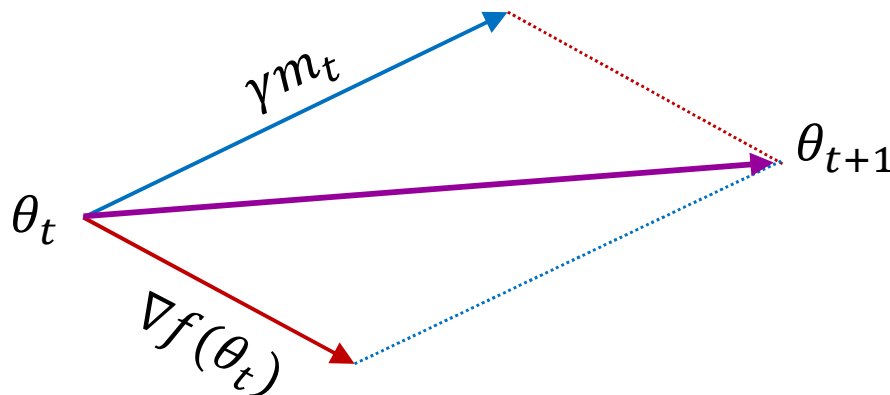
- Dao động gần cực tiểu khi hệ số quán tính quá lớn.



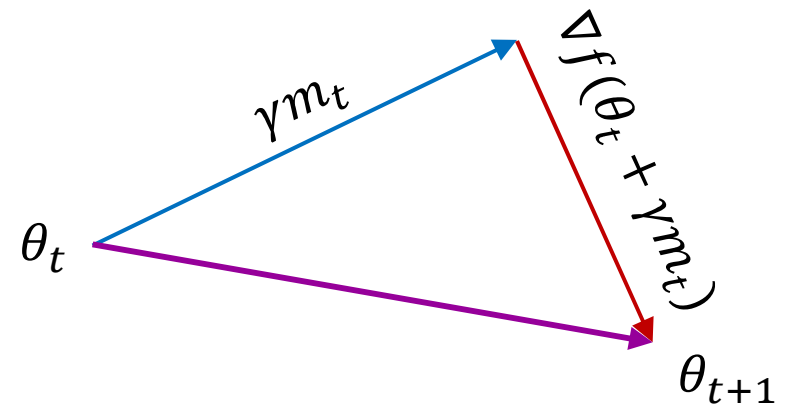
Giới thiệu đề tài

Đề tài liên quan: NAG

- Cải tiến: [Nesterov Accelerated Descent](#) (NAG)*
 - Tính đạo hàm tại **(điểm hiện tại + quán tính)** để lấy hướng cập nhật tiếp theo rồi mới cộng quán tính vào lượng cập nhật.



Momentum



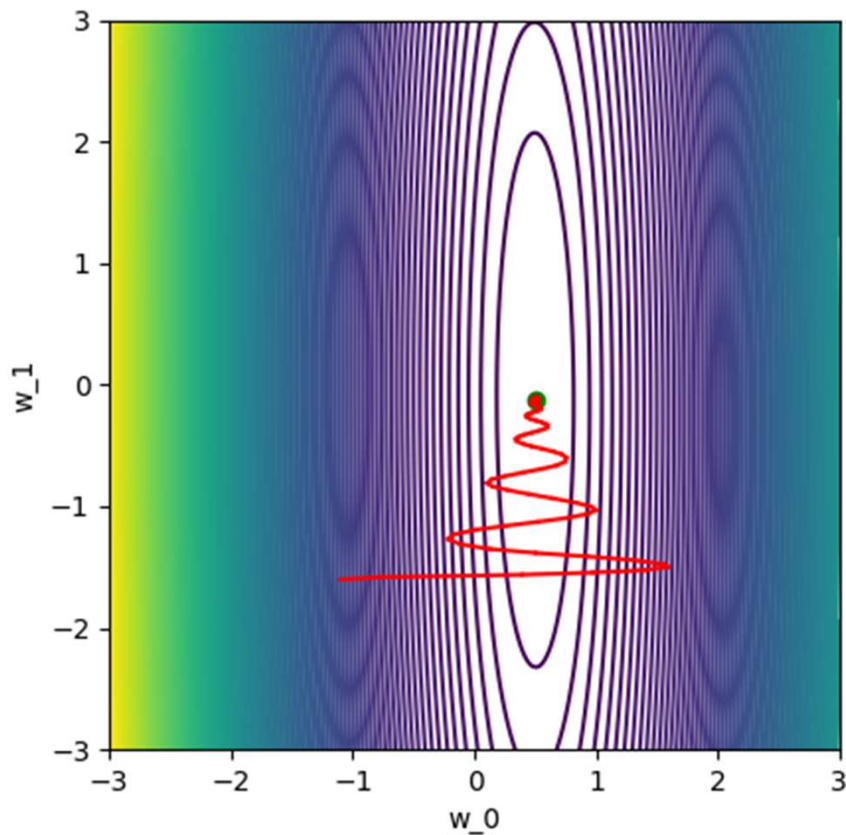
Nesterov

*Nesterov, Y., "A method of solving a convex programming problem with convergence rate $O(1/\sqrt{k})$ ", *Soviet Mathematics Doklady*, 1983, vol. 27, pp. 372-376.

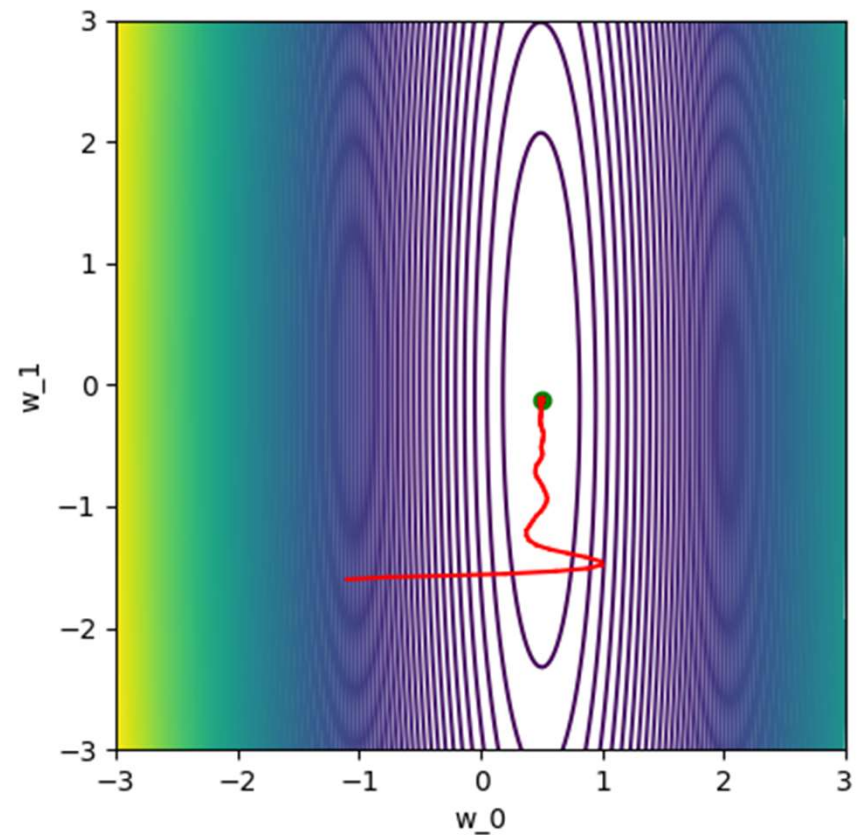
Giới thiệu đề tài

Đề tài liên quan: NAG

- Đường đi ổn định hơn.



SGD with Momentum

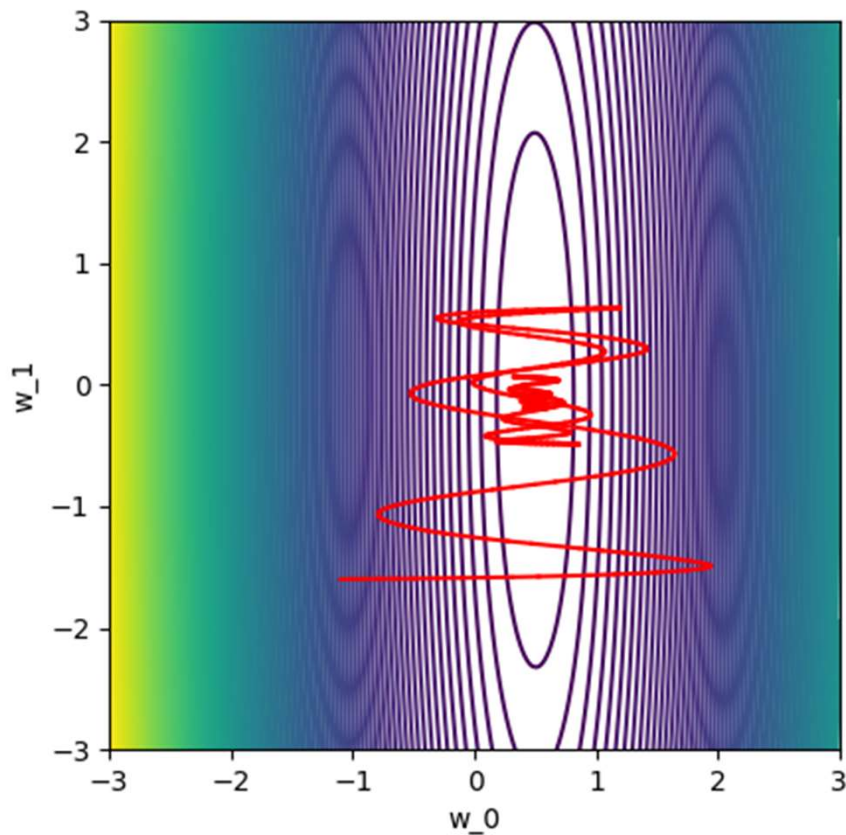


SGD with Nesterov

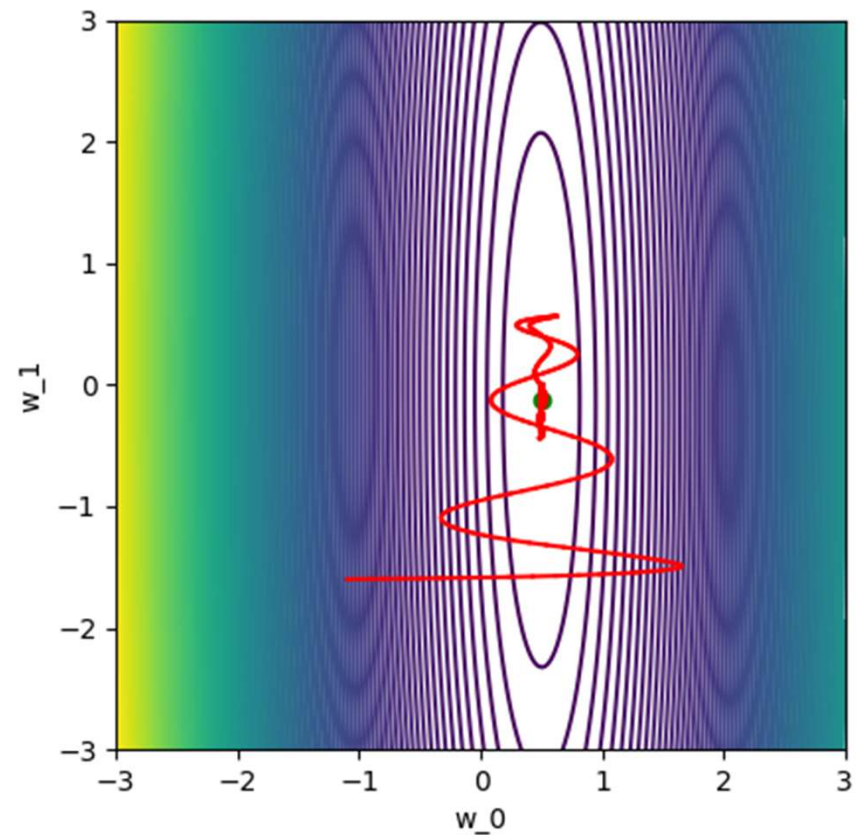
Giới thiệu đề tài

Đề tài liên quan: NAG

- Cho phép sử dụng hệ số momentum lớn hơn.



SGD with Momentum

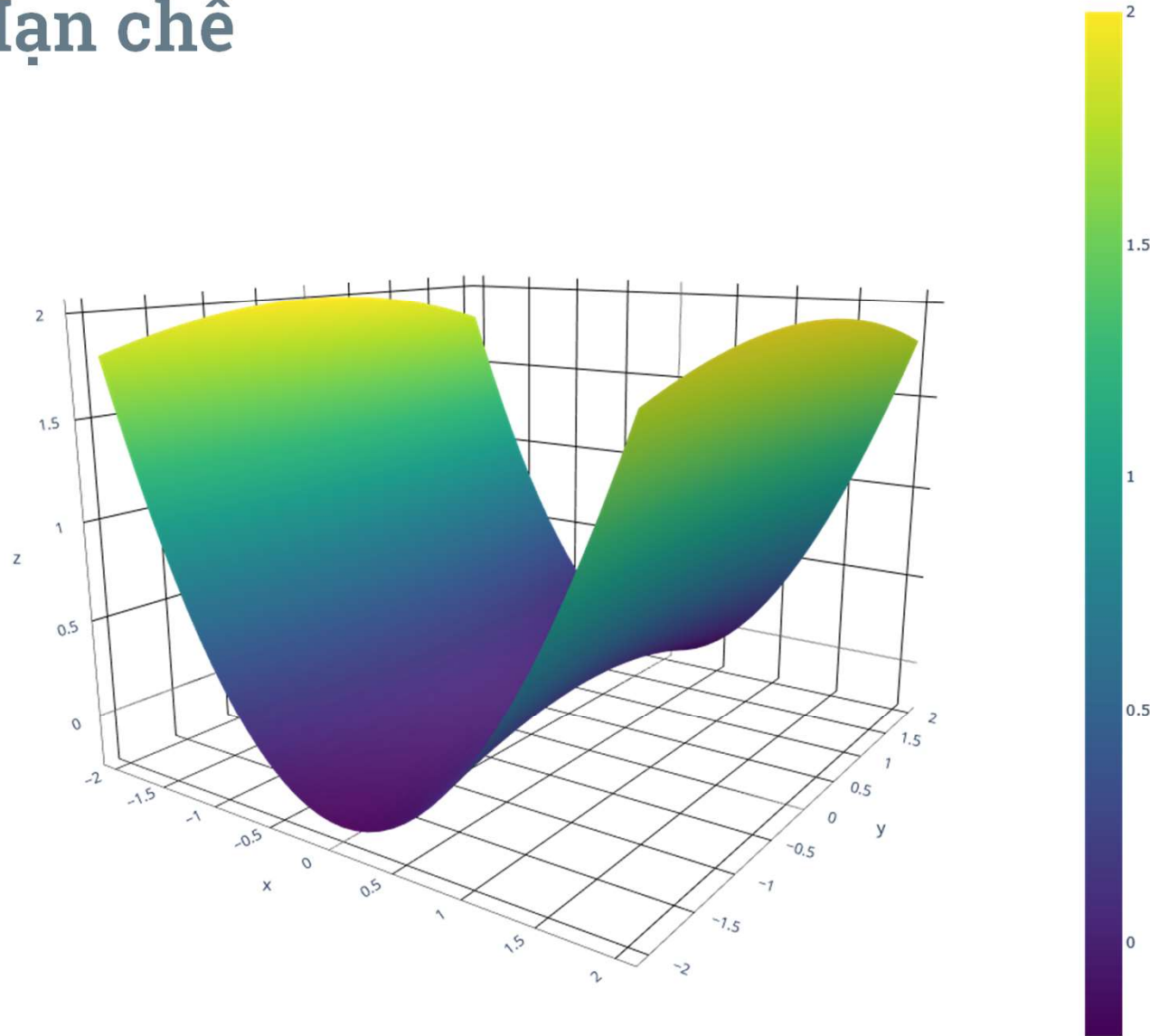


SGD with Nesterov

Giới thiệu đề tài

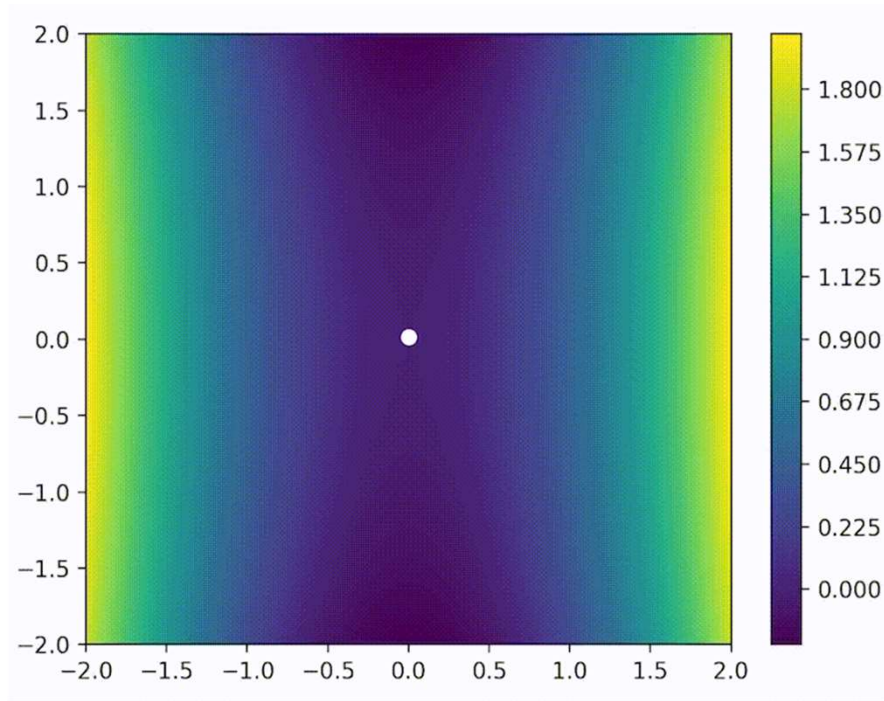
Đề tài liên quan: Hạn chế

- Tỷ lệ học cố định không phù hợp cho tất cả các hướng.

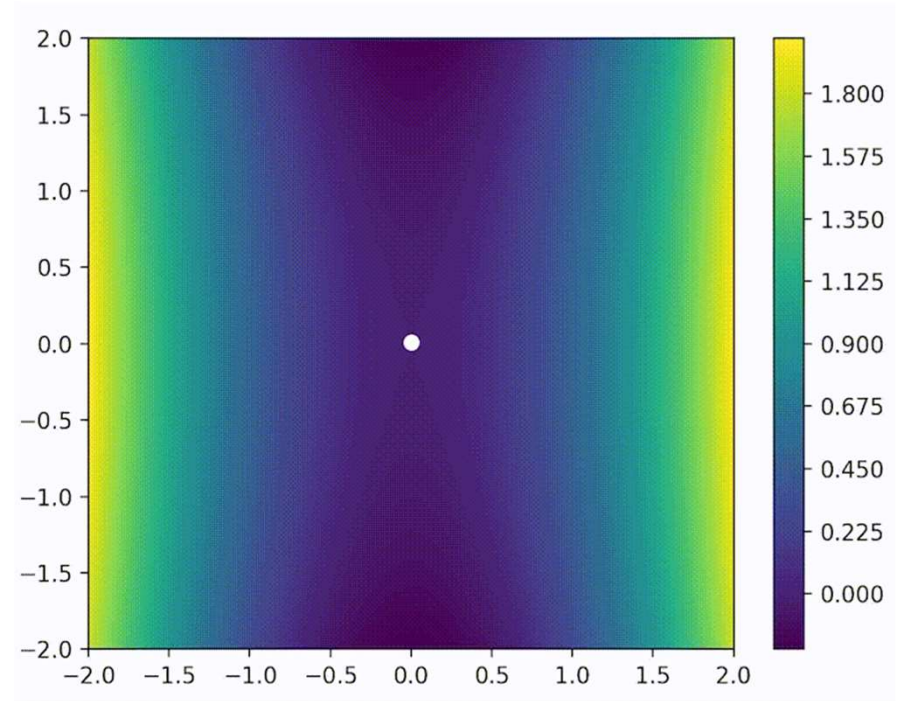


Giới thiệu đề tài

Đề tài liên quan: Hạn chế



Adam



GD with Momentum

Giới thiệu đề tài

Đề tài liên quan: Hạn chế

- Cần một tỉ lệ học phù hợp cho từng hướng.
- Adaptive learning rate

Giới thiệu đề tài

Bài báo tìm hiểu

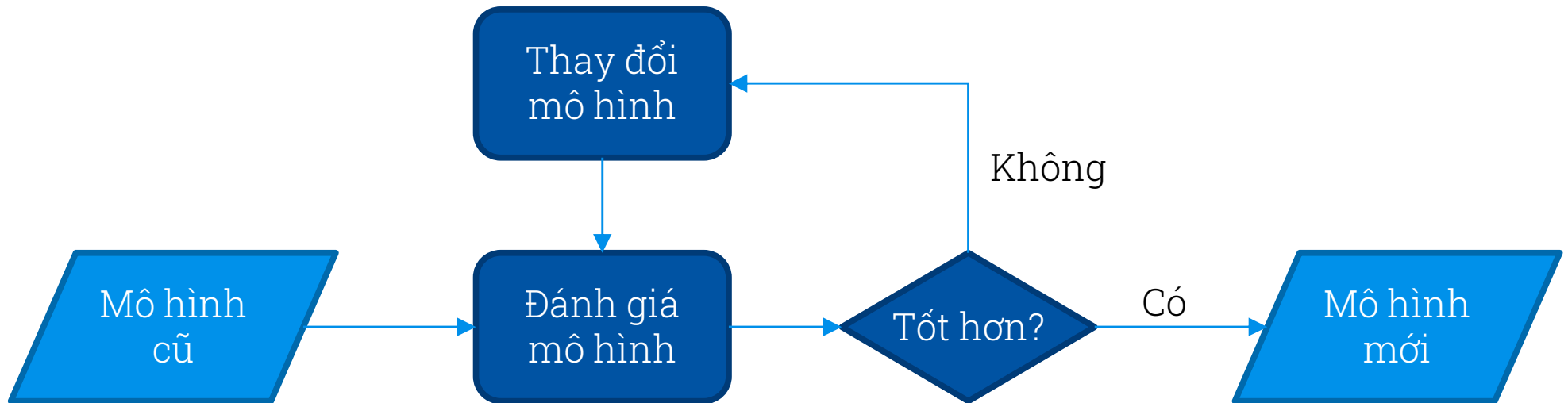
- "Adam: A method for stochastic optimization",
Diederik P. Kingma, Jimmy Lei Ba (2014).
 - Sử dụng tỉ lệ học riêng biệt cho từng trọng số.
 - Hội tụ về điểm cực tiểu có độ lỗi thấp.

2.

Kiến thức nền tảng

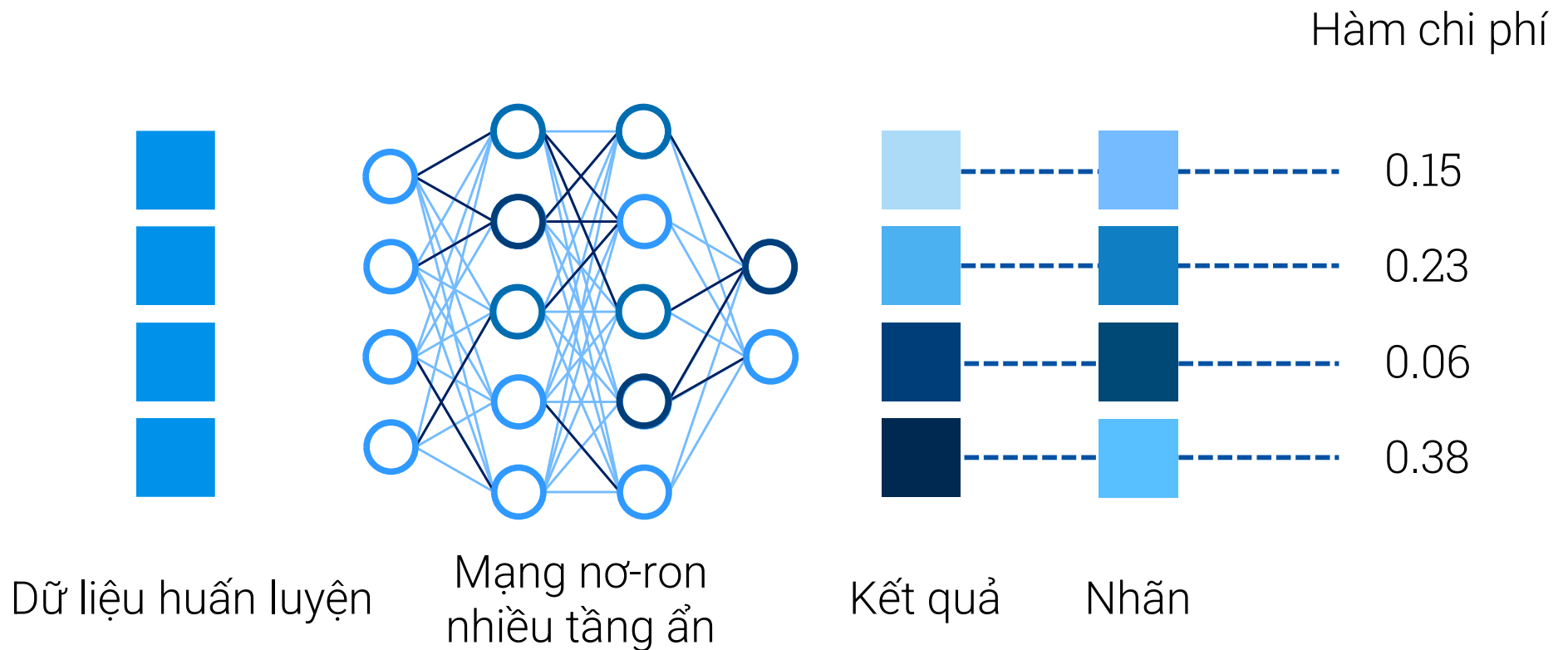
Kiến thức nền tảng

Tối ưu hóa (optimization)



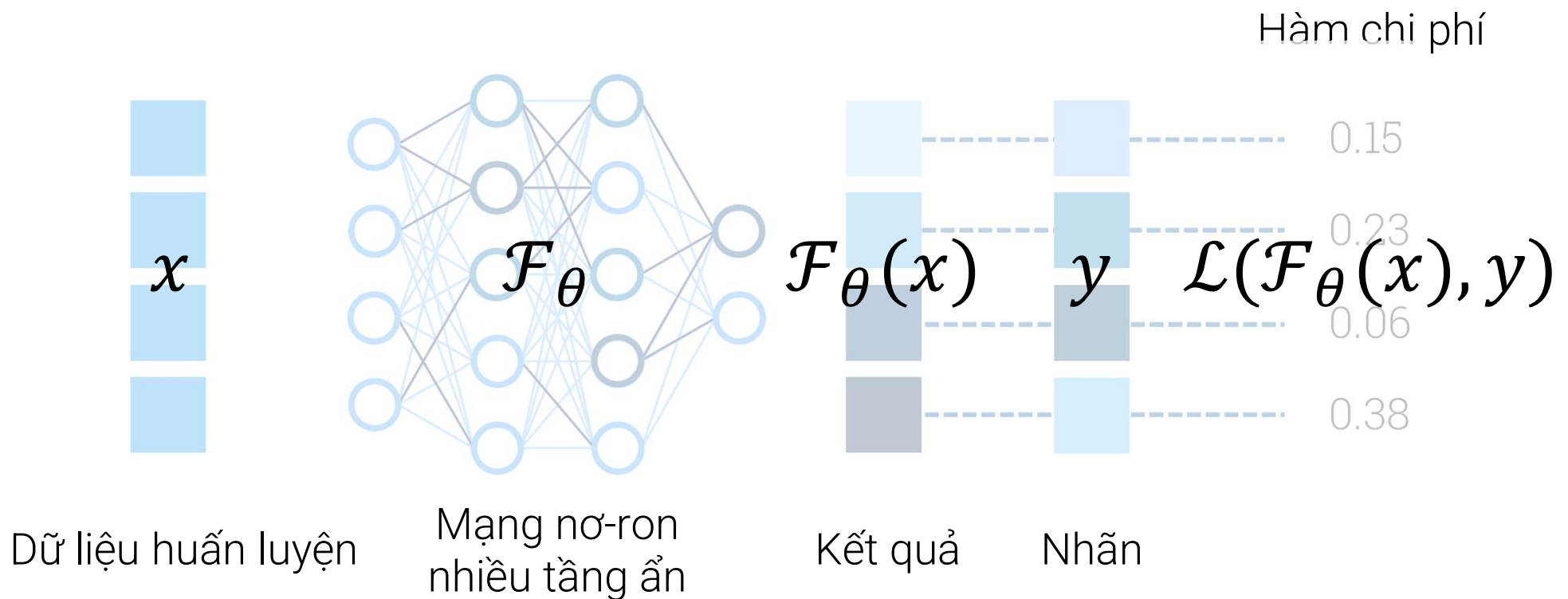
Kiến trúc nền tảng

Tối ưu hóa (optimization)



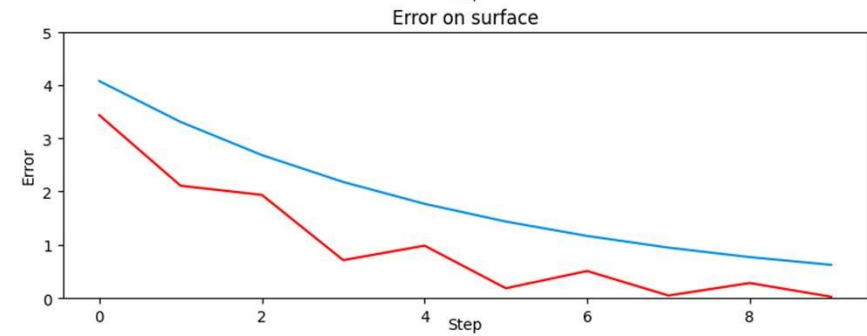
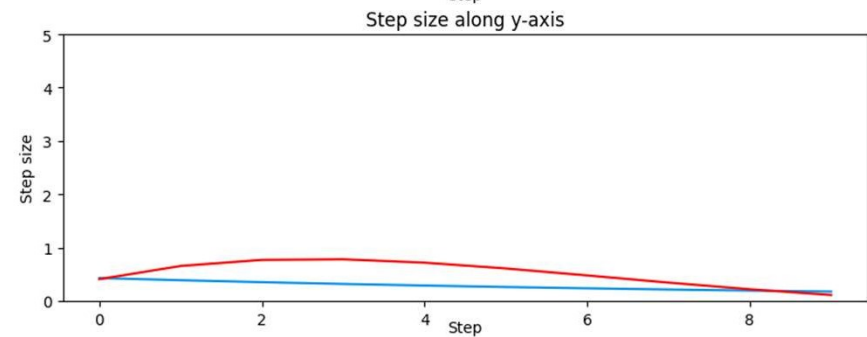
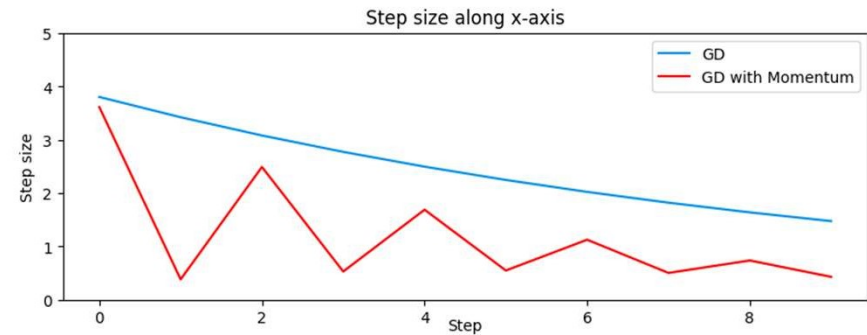
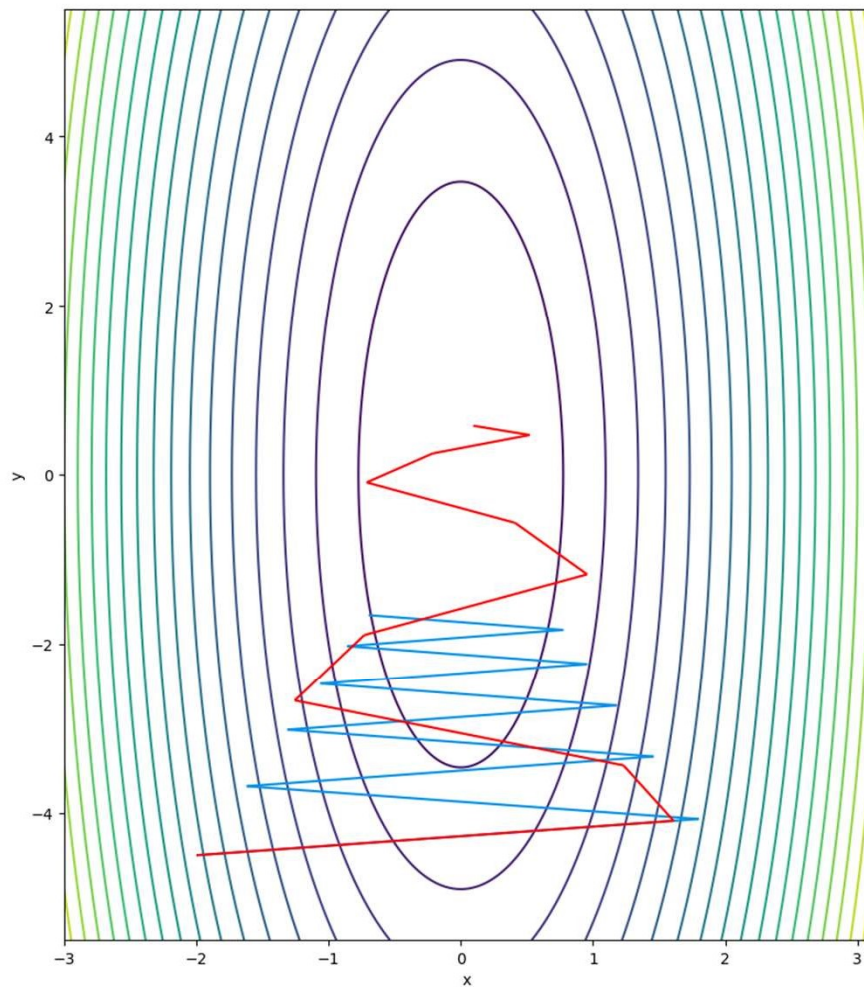
Kiến trúc nền tảng

Tối ưu hóa (optimization)



Kiến thức nền tảng

Momentum



Kiến thức nền tảng

AdaGrad

- Tính đạo hàm theo từng tham số.

$$g_t = \nabla_{\theta} \mathcal{L}(\theta_t)$$

- Cập nhật G ở bước hiện tại.

$$G_t = G_{t-1} + \text{diag}(g_t \cdot g_t^T)$$

- Cập nhật trọng số.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

Kiến trúc nền tảng

AdaGrad

$$g_t = \begin{bmatrix} g_{t,1} \\ g_{t,2} \\ g_{t,3} \\ \vdots \\ g_{t,n} \end{bmatrix} \quad g_t^T = [g_{t,1} \quad \dots \quad g_{t,n}]$$

$$g_t \cdot g_t^T = \begin{bmatrix} g_{t,1} \\ g_{t,2} \\ g_{t,3} \\ \vdots \\ g_{t,n} \end{bmatrix} [g_{t,1} \quad \dots \quad g_{t,n}] = \begin{bmatrix} \mathbf{g}_{t,1}^2 & \dots & g_{t,1} \cdot g_{t,n} \\ \vdots & \ddots & \vdots \\ g_{t,n} \cdot g_{t,1} & \dots & \mathbf{g}_{t,n}^2 \end{bmatrix}$$

Kiến thức nền tảng

AdaGrad

- Tính đạo hàm theo từng tham số.

$$g_t = \nabla_{\theta} \mathcal{L}(\theta_t)$$

- Cập nhật G ở bước hiện tại.

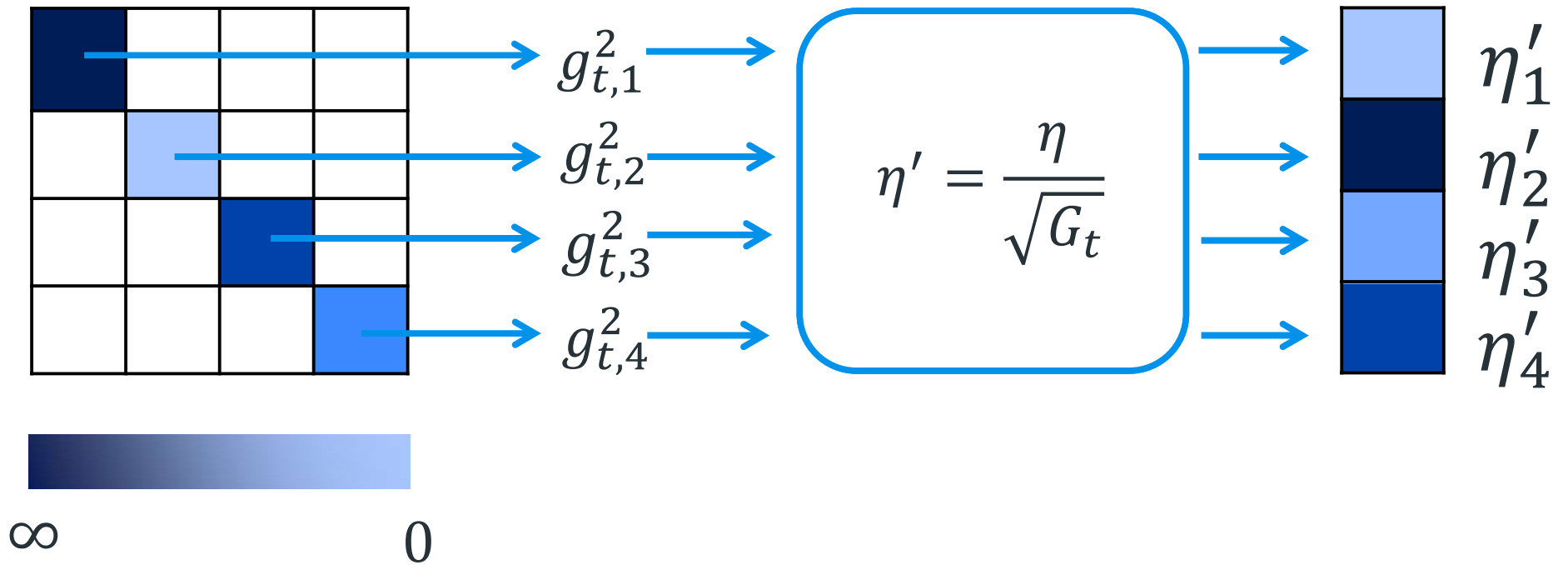
$$G_t = G_{t-1} + \mathbf{g}_t^2$$

- Cập nhật trọng số.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

Kiến trúc nền tảng

AdaGrad



Kiến thức nền tảng

AdaGrad và RMSprop

AdaGrad

- $g_t = \nabla_{\theta} \mathcal{L}(\theta_t)$
- $G_t = G_{t-1} + g_t^2$
- $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$

giá trị không âm

RMSprop

- $g_t = \nabla_{\theta} \mathcal{L}(\theta_t)$
- $G_t = \gamma G_{t-1} + (1 - \gamma) g_t^2$
- $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$

"leaky"

3.

Thuật toán Adam

Thuật toán Adam

Siêu tham số

- α : tỉ lệ học
- β_1, β_2 : tỉ lệ suy biến của trung bình đạo hàm và bình phương đạo hàm (mặc định lần lượt là 0.9 và 0.999)
- ϵ : hệ số nhỏ

Thuật toán Adam

Các bước thực hiện

Tăng bước chạy t

$$t = t + 1$$

Tính đạo hàm của hàm chi phí trên từng tham số

$$g_t = \nabla_{\theta} f_t(\theta_{t-1})$$

Thuật toán Adam

Các bước thực hiện

Cập nhật \mathbf{m}_t và \mathbf{v}_t

$$\begin{aligned} m_t &= \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t & \text{--- Momentum} \\ v_t &= \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 & \text{--- RMSprop} \end{aligned}$$

Tính **bias-correction** của m_t và v_t

$$\begin{aligned} \hat{m}_t &= m_t / (1 - \beta_1^t) \\ \hat{v}_t &= v_t / (1 - \beta_2^t) \end{aligned}$$

t tăng dần $\rightarrow \beta^t$ giảm dần $\rightarrow 1 - \beta^t$ tiến dần đến 1

Cập nhật **trọng số**

$$\theta_t = \theta_{t-1} - \alpha \cdot \hat{m}_t / \sqrt{\hat{v}_t} + \epsilon$$

4.

Thí nghiệm

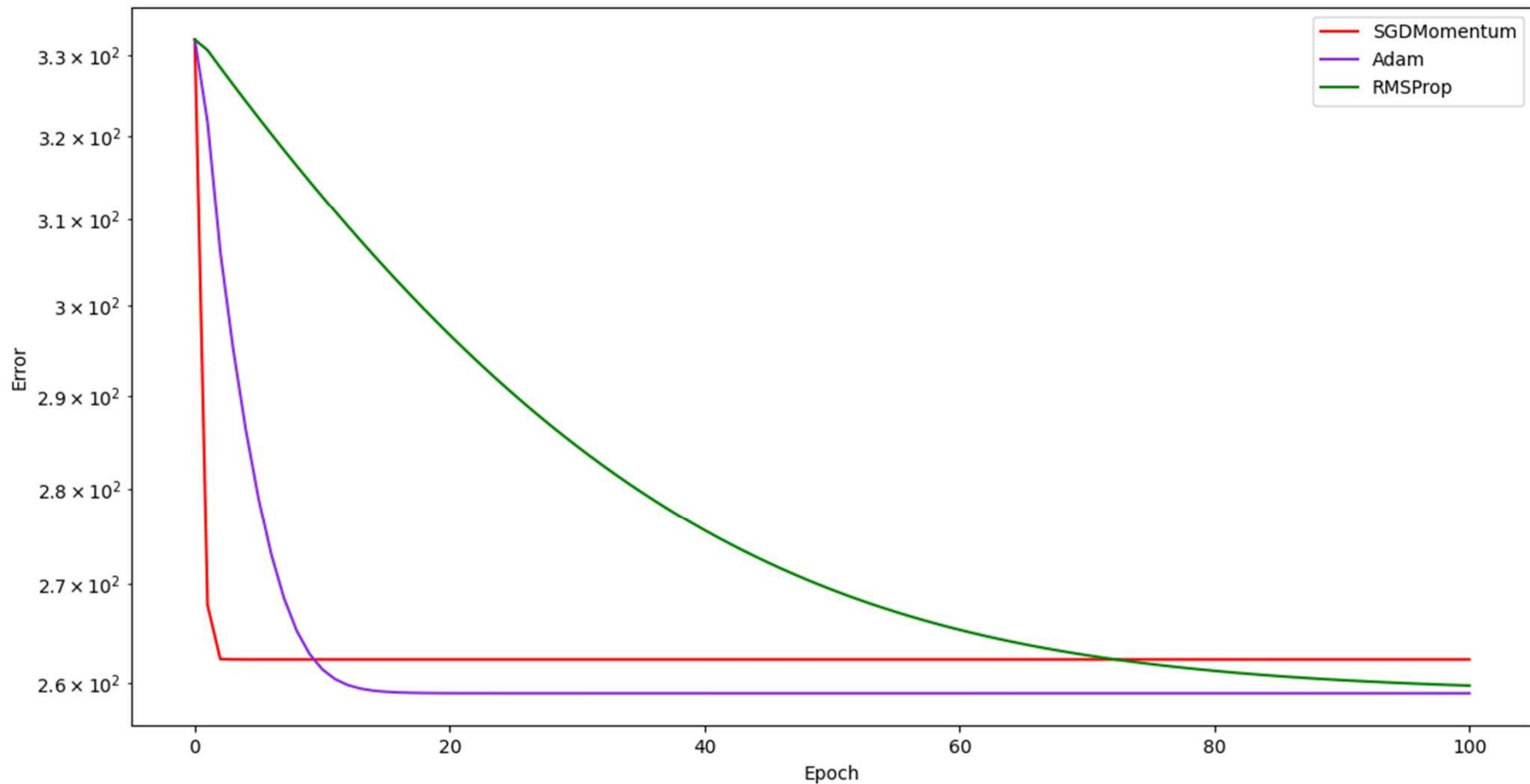
Thí nghiệm

Cách thực hiện

- Khởi tạo mạng nơ-ron với các tham số ngẫu nhiên.
- Lưu các tham số này làm điểm xuất phát chung.
- Với mỗi thuật toán:
 - Nạp lại bộ trọng số đã lưu ở trên cho mạng nơ-ron.
 - Thực hiện tối ưu hóa mạng nơ-ron với số bước xác định.
 - Ghi nhận độ lỗi tại mỗi epoch.

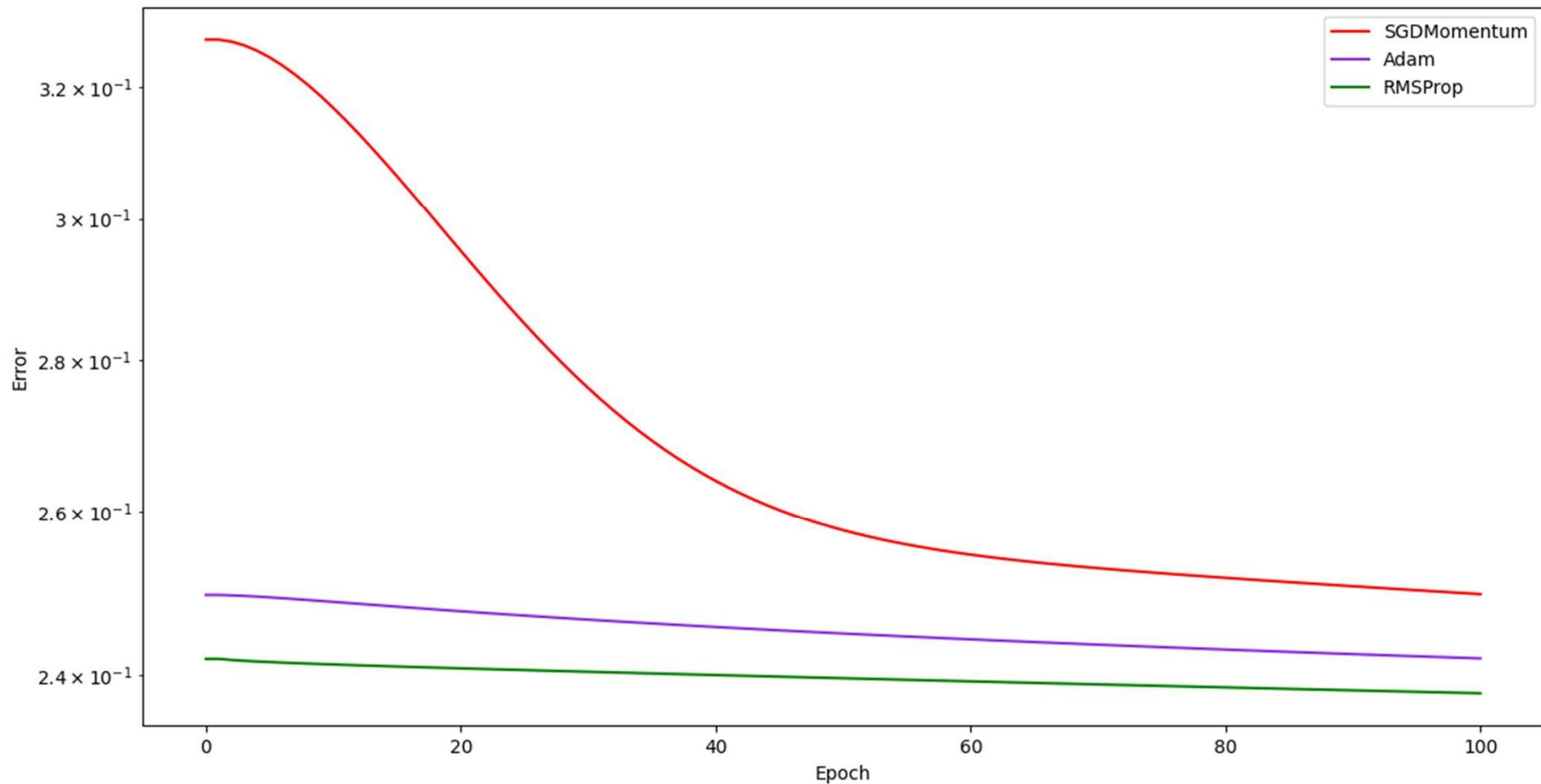
Thí nghiệm

Linear regression với độ nhiễu cao



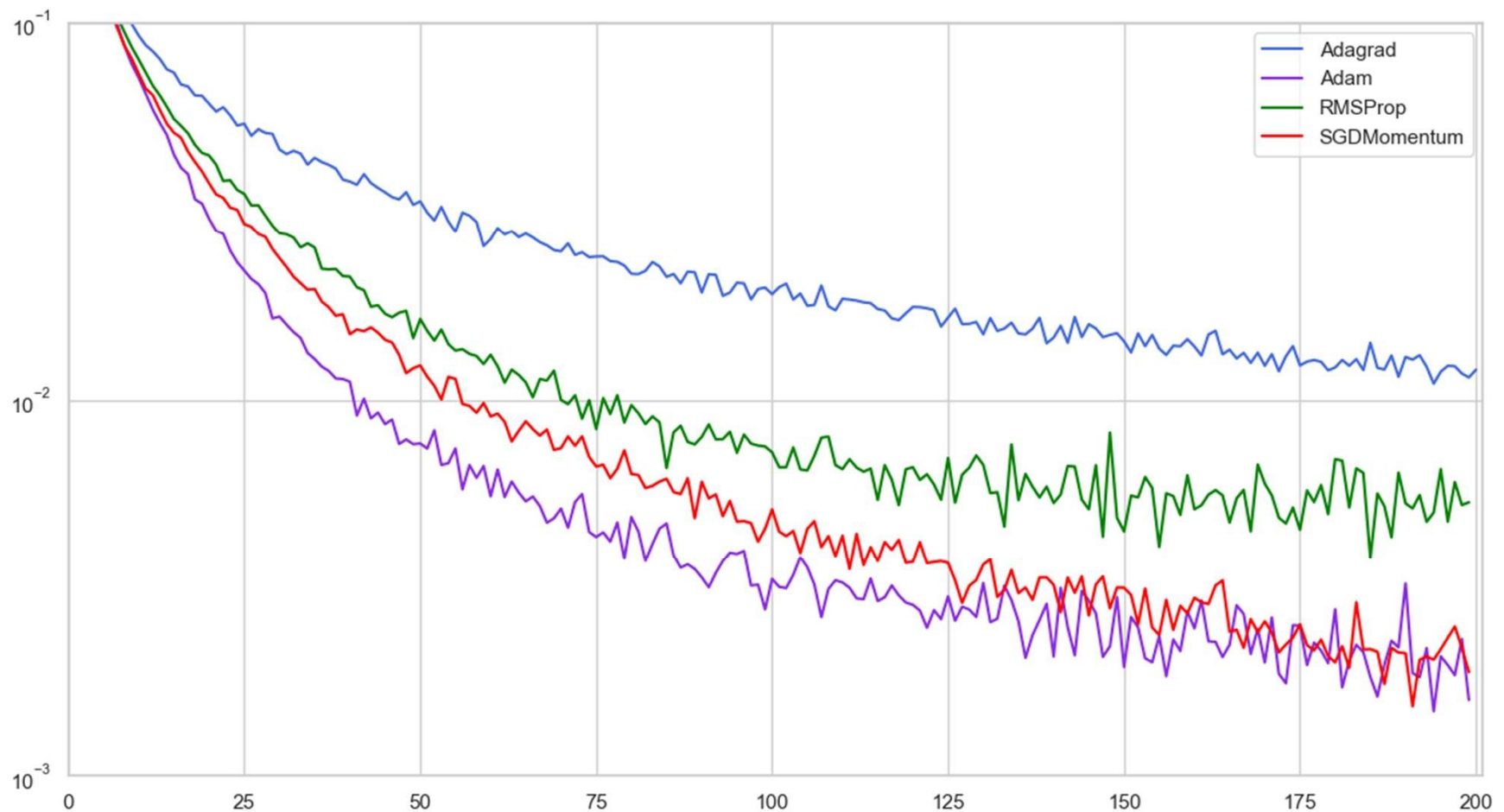
Thí nghiệm

Classification với dữ liệu thưa



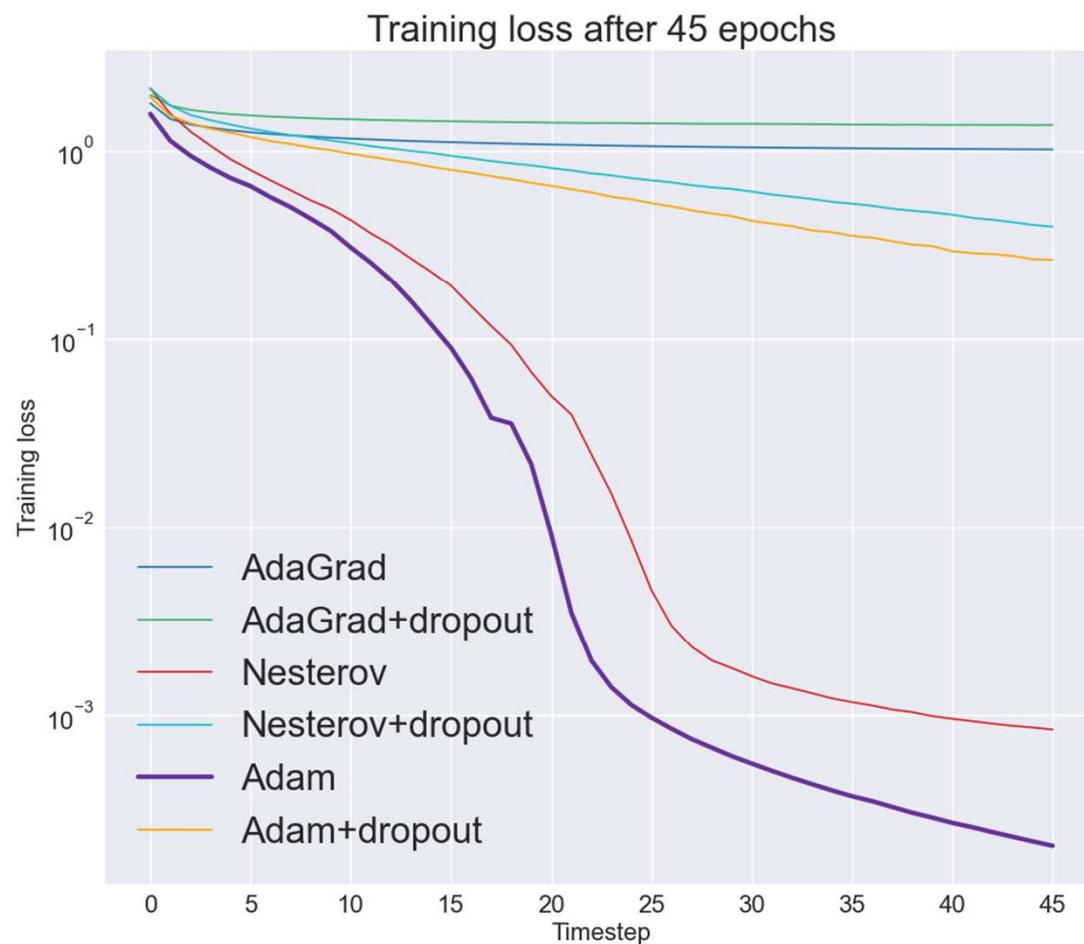
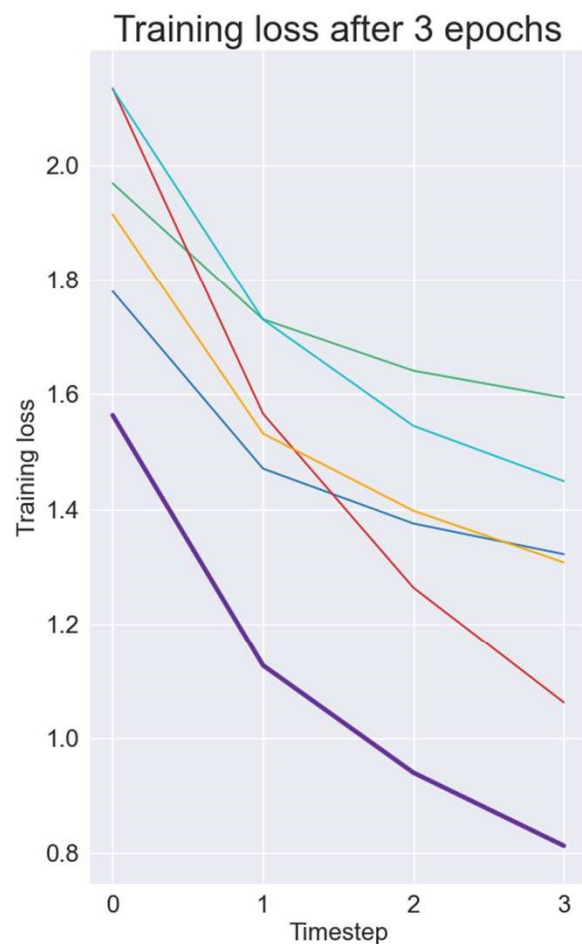
Thí nghiệm

1000fc-1000fc - MNIST



Thí nghiệm

c64-c64-c128-1000FC – CIFAR10



Thí nghiệm

Siêu tham số

	Learning Rate	Momentum/ Beta 1	Alpha/ Beta 2	Epsilon
SGDMomentum	0.01	0.9	-	10^{-8}
AdaGrad	0.01	-	-	
RMSprop	0.0001	-	0.9	
Adam	0.0001	0.9	0.999	

5.

Tổng kết

Tài liệu tham khảo

- Yann Dauphin *et al.*, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization", in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2933-2941.
- Ning Qian, "On the momentum term in gradient descent learning algorithms", *Journal of the International Neural Network Society*, 1999, vol. 12, pp. 145-151.
- Nesterov, Y., "A method of solving a convex programming problem with convergence rate $O(1/\sqrt{k})$ ", *Soviet Mathematics Doklady*, 1983, vol. 27, pp. 372-376.
- J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, 2011.

Tài liệu tham khảo

- T. Tieleman and G. Hinton, “Lecture 6.5 - rmsprop,” *COURSERA: Neural Networks for Machine Learning*, 2012.
- D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference for Learning Representation*, vol. abs/1412.6980, 2015.