

AI VIETNAM
All-in-One Course

Optimization Algorithms in Deep Learning

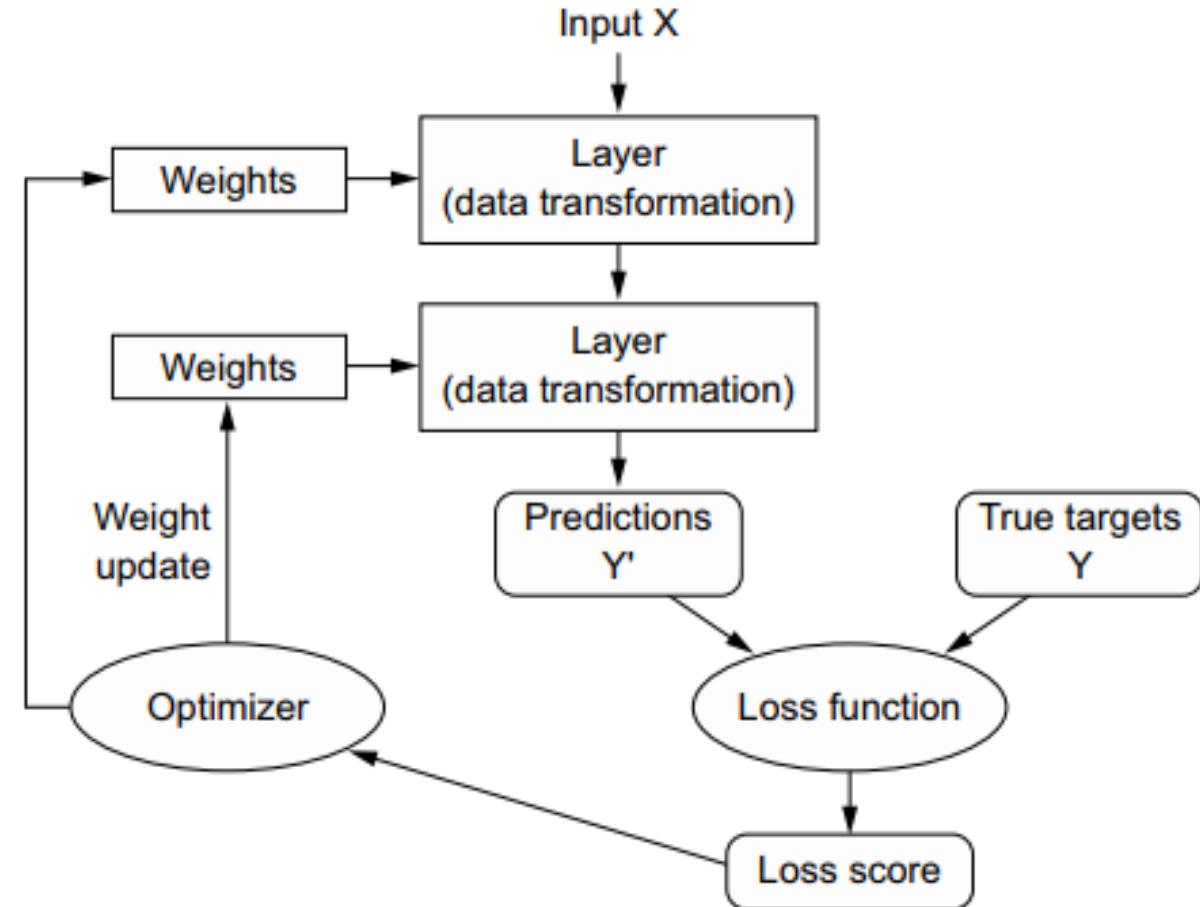
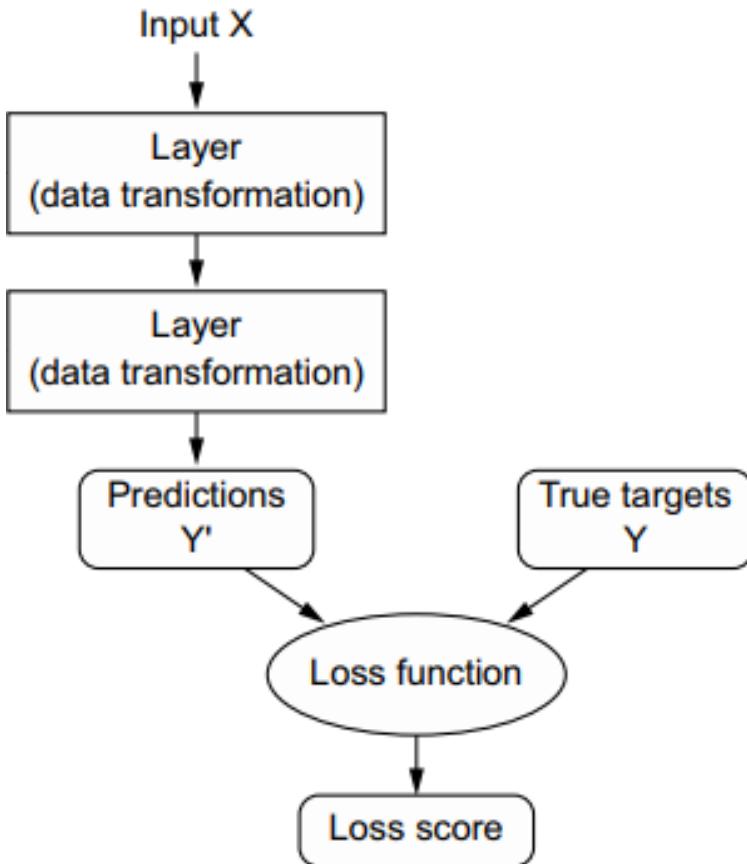
Quang-Vinh Dinh
PhD in Computer Science

Outline

- Quick Introduction
- Stochastic Gradient Descent
- Adagrad
- RMSProp
- Momentum
- Adam
- Case studies

Optimization Algorithms

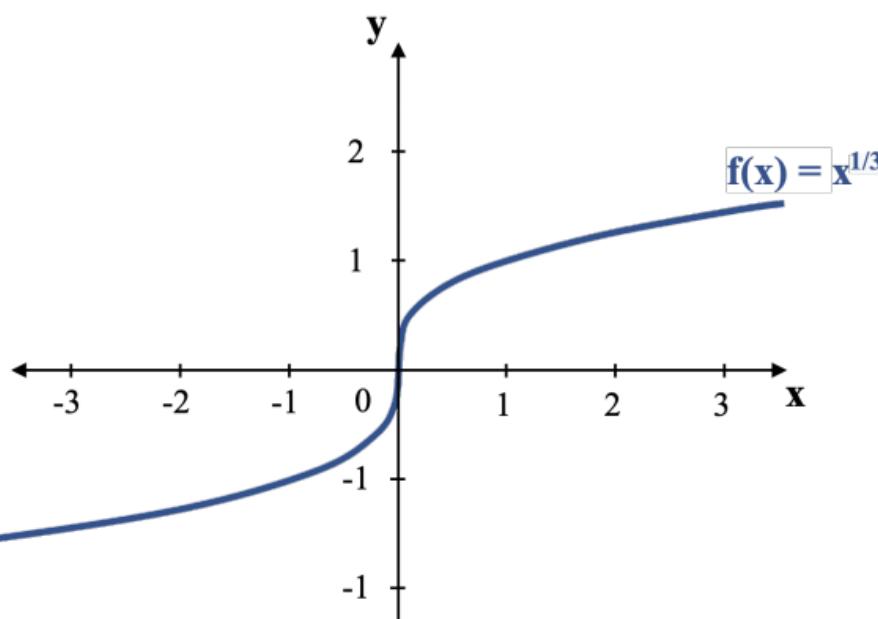
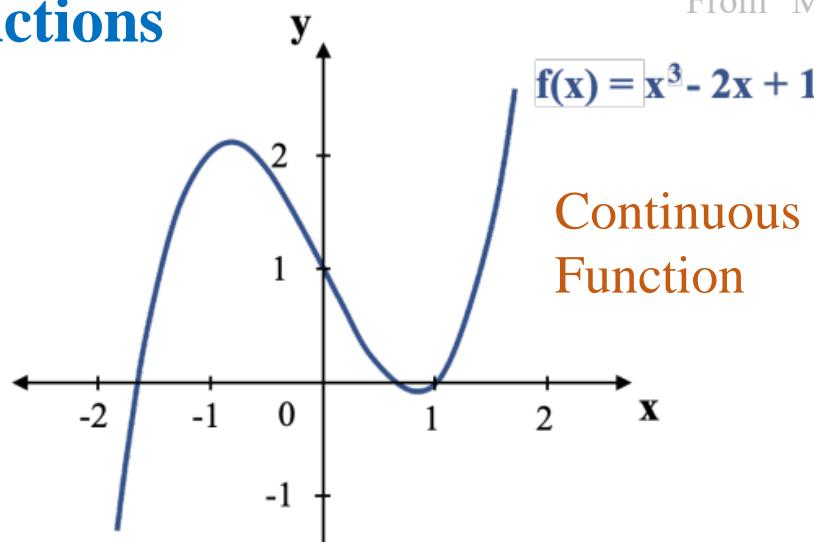
Overview



Optimization Algorithms

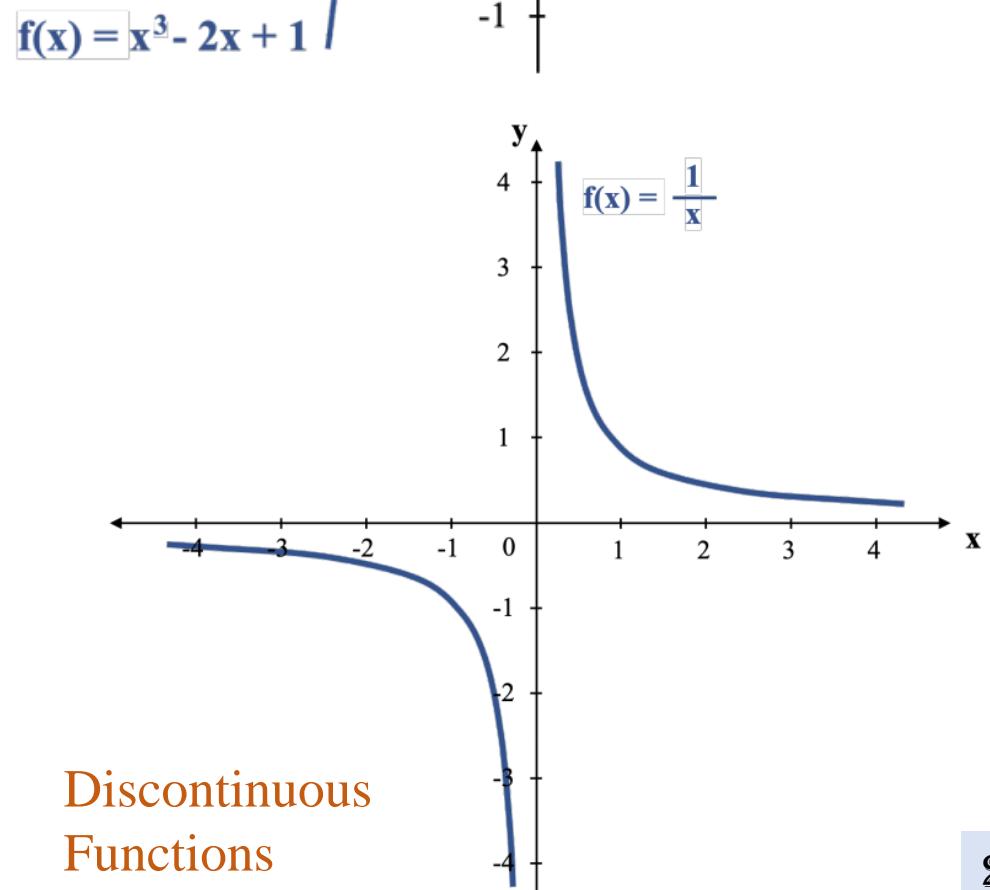
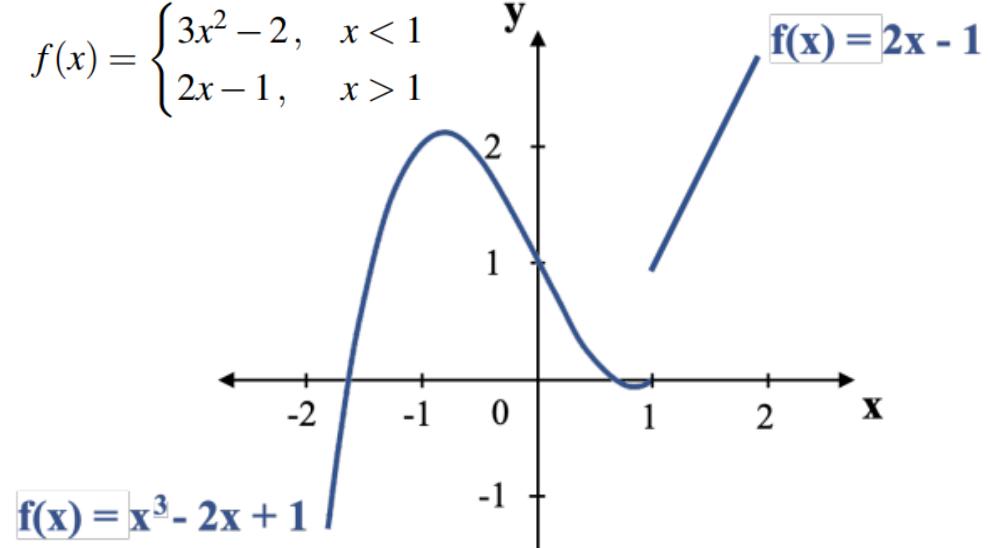
Loss functions

From "Machine Learning Simplified"



Continuous non-differentiable functions

$$f(x) = \begin{cases} 3x^2 - 2, & x < 1 \\ 2x - 1, & x > 1 \end{cases}$$



Discontinuous Functions

Optimization Algorithms

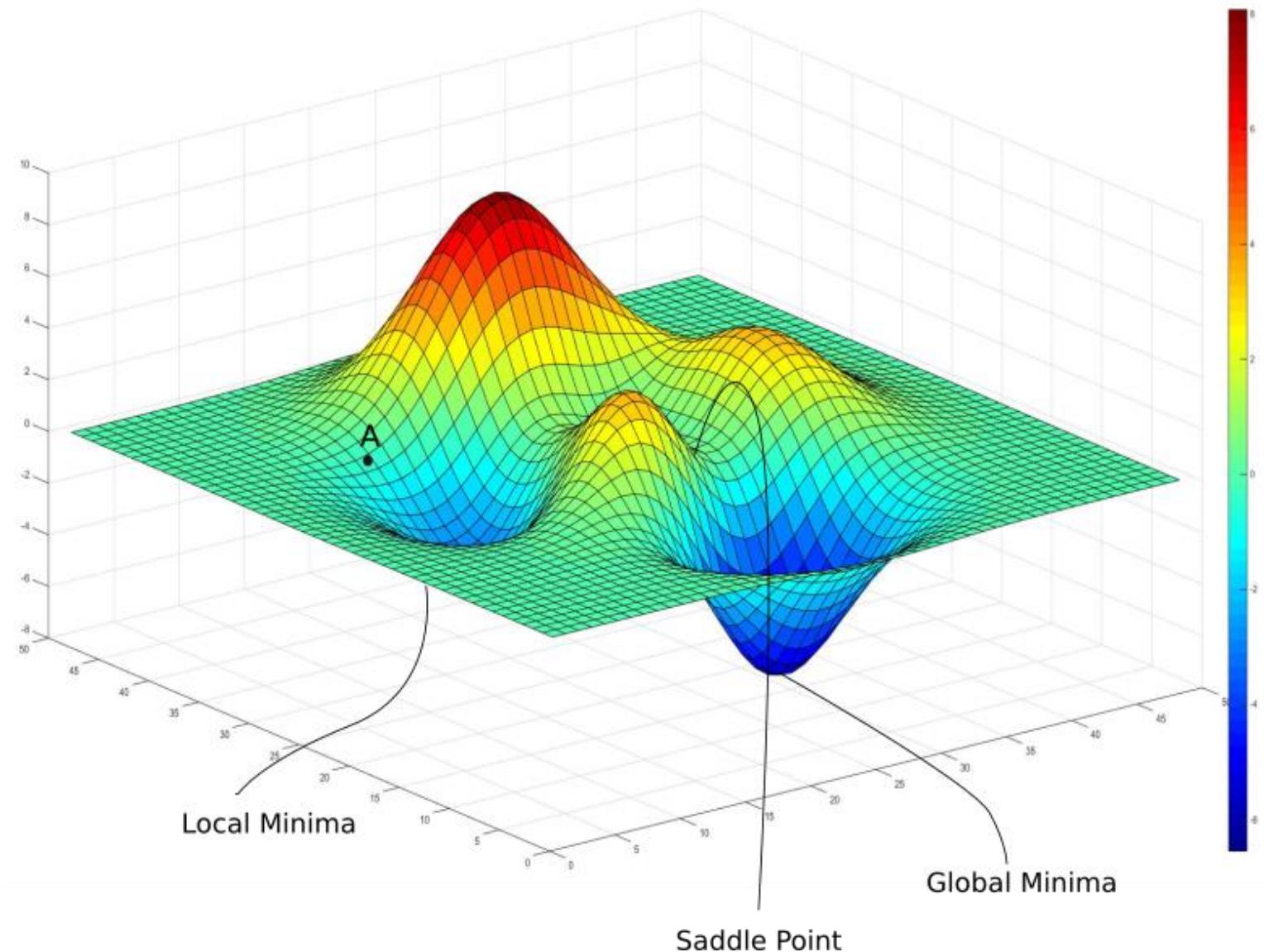
Challenges

Local minima

Global minima

Saddle points

<https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/>



Optimization Algorithms

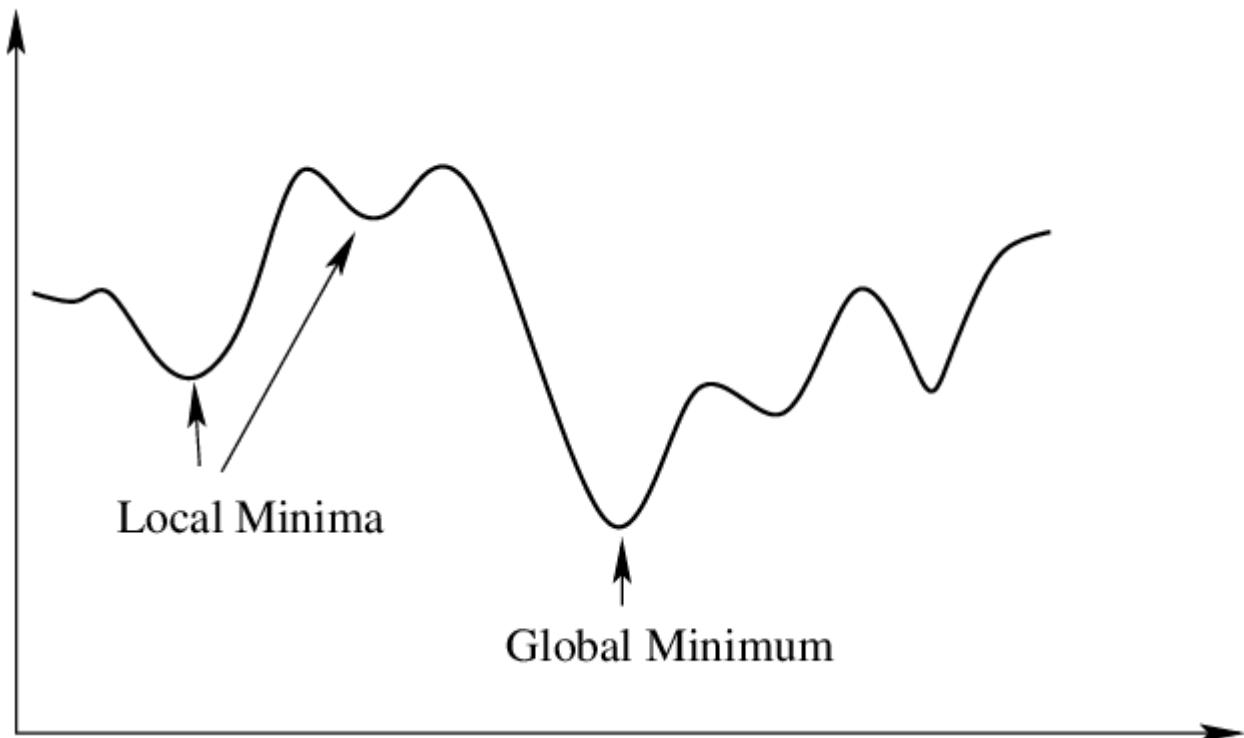
Challenges

Local minima

Global minima

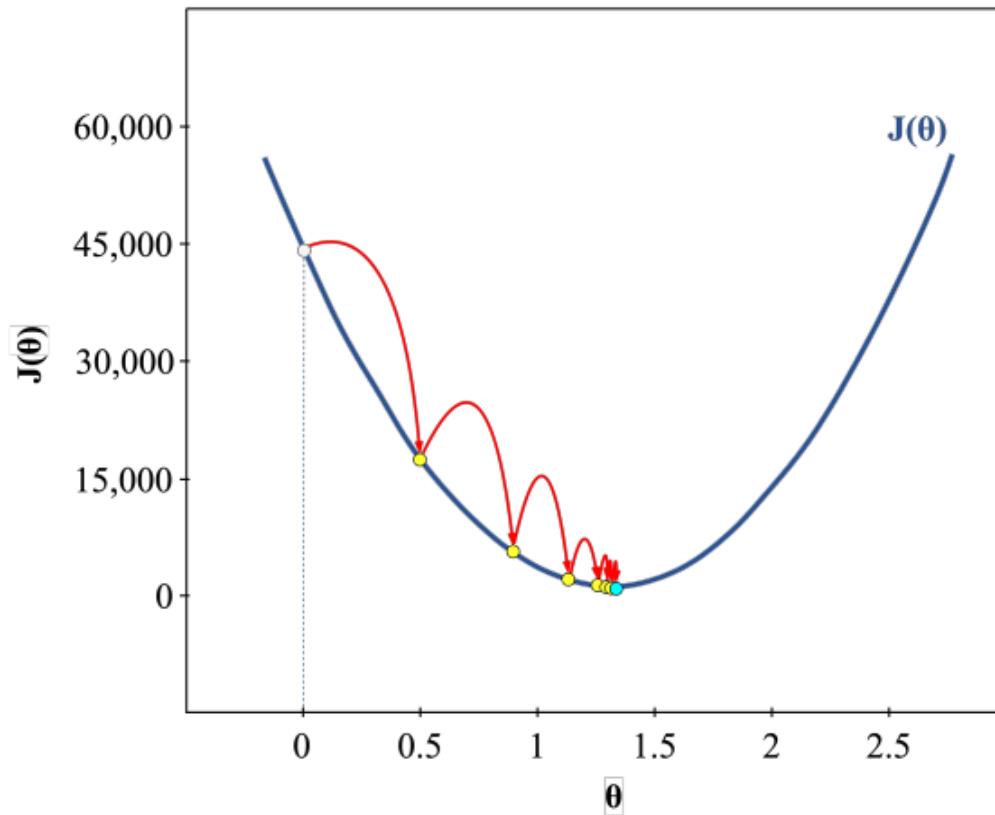
Saddle points

<https://vitalflux.com/local-global-maxima-minima-explained-examples/>

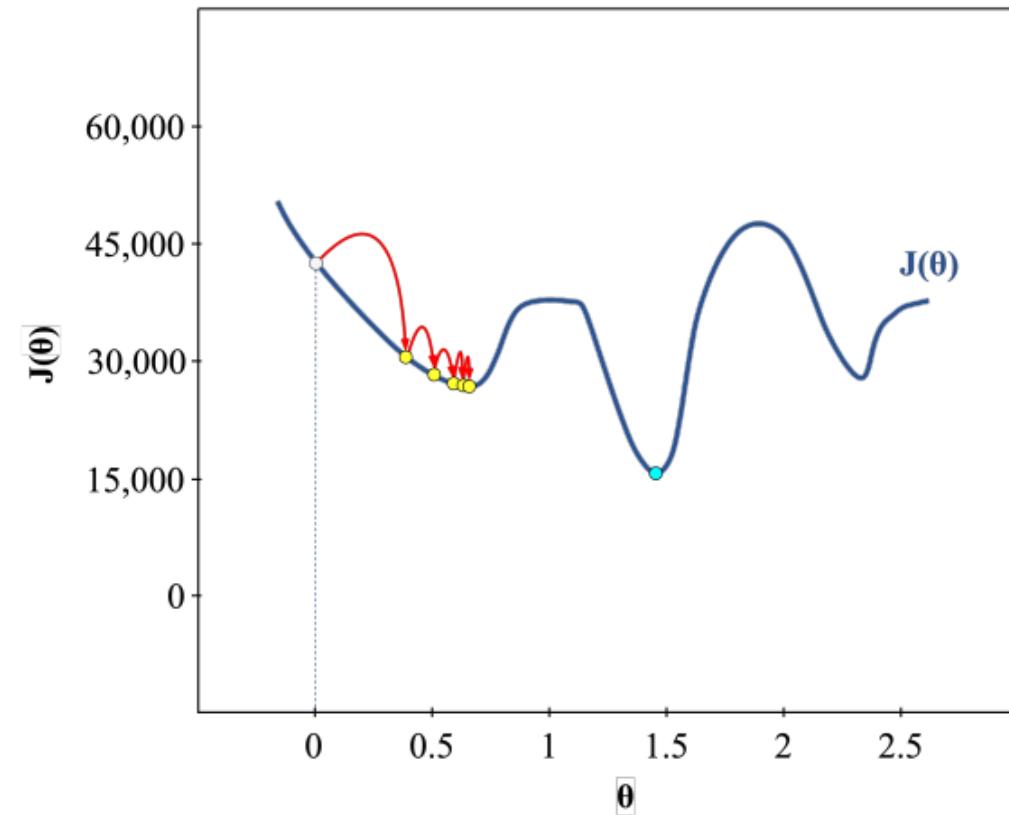


Optimization Algorithms

Challenges: Local minima



(a) Gradient Descent on a Convex Cost Function.

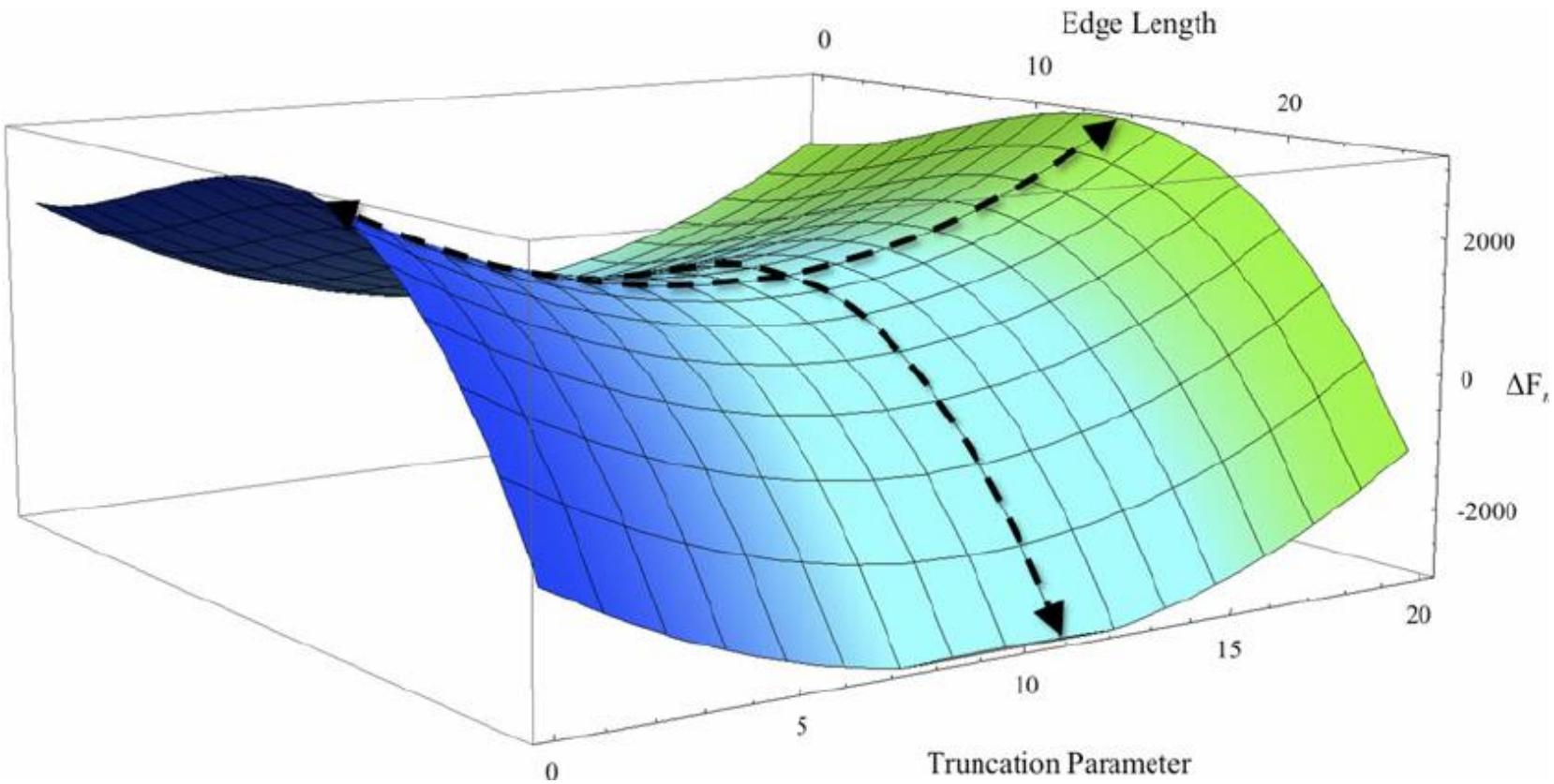
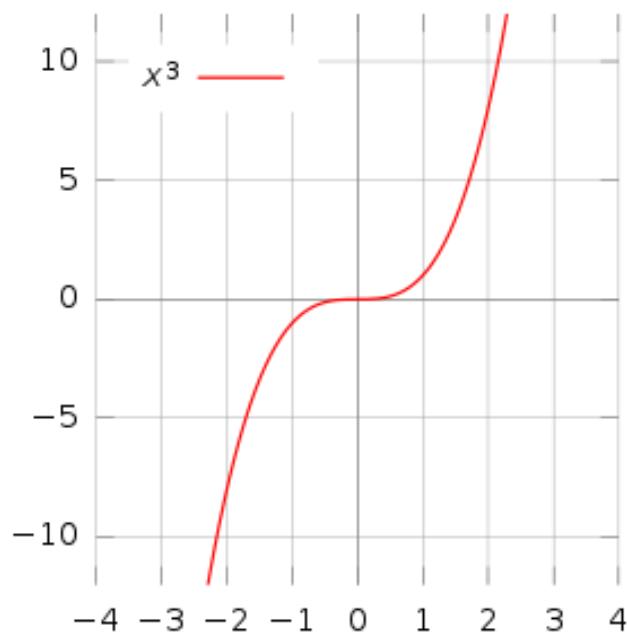


(b) Gradient Descent on a Non-convex Cost Function.

Optimization Algorithms

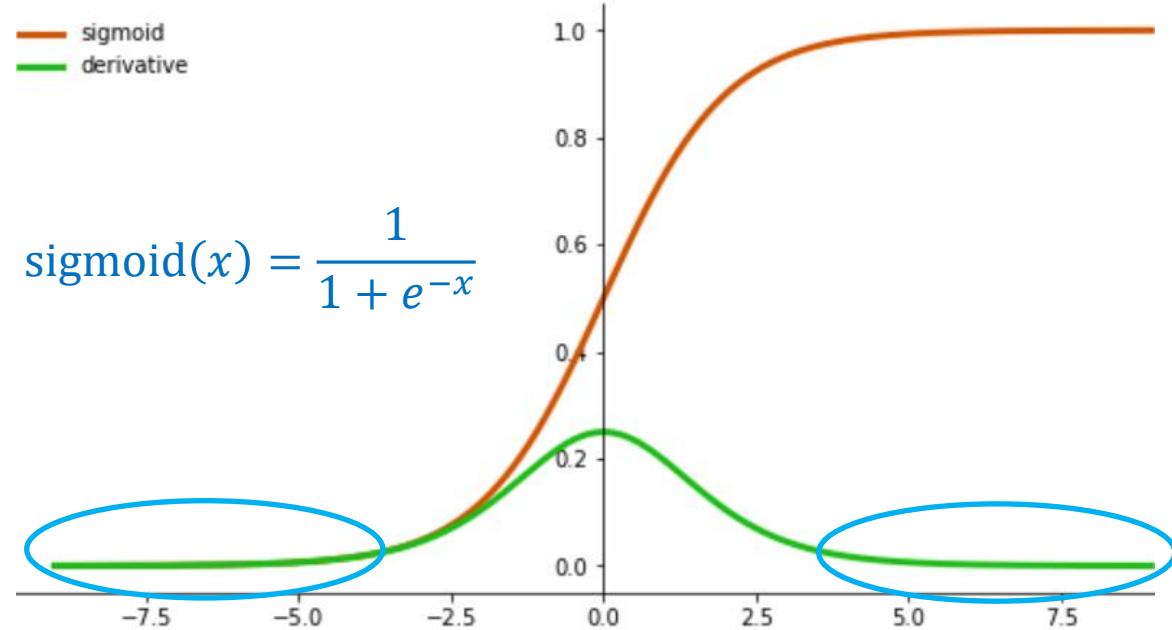
Challenges

Saddle points

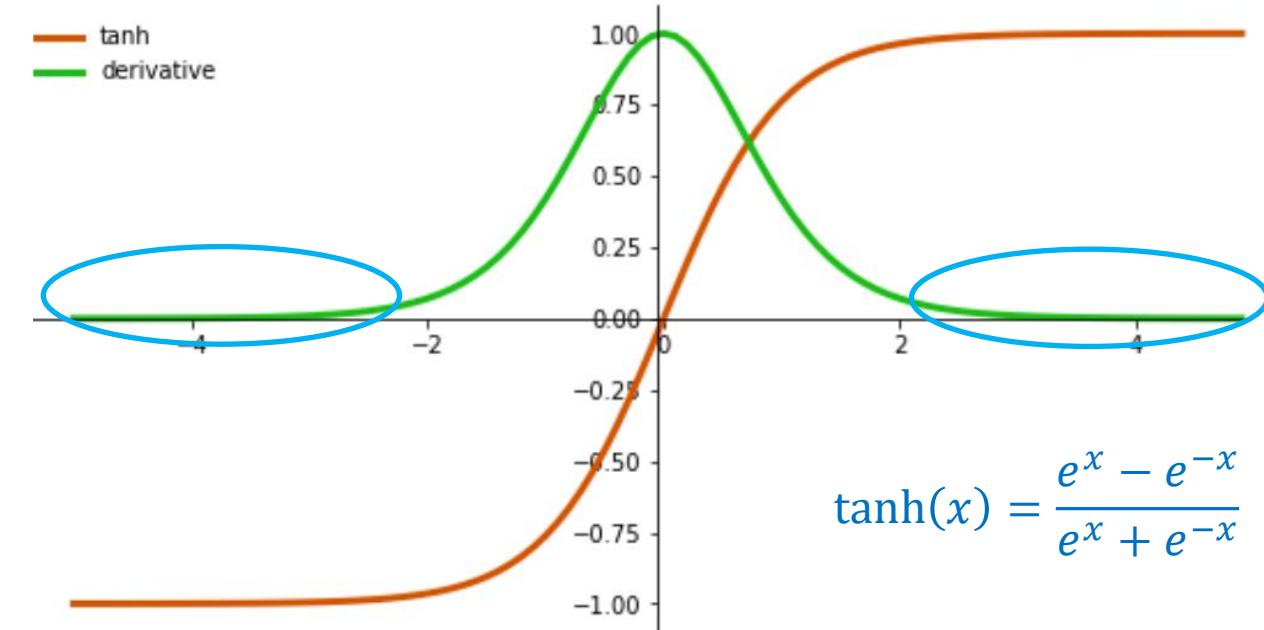


Optimization Algorithms

Challenges: Gradient vanishing



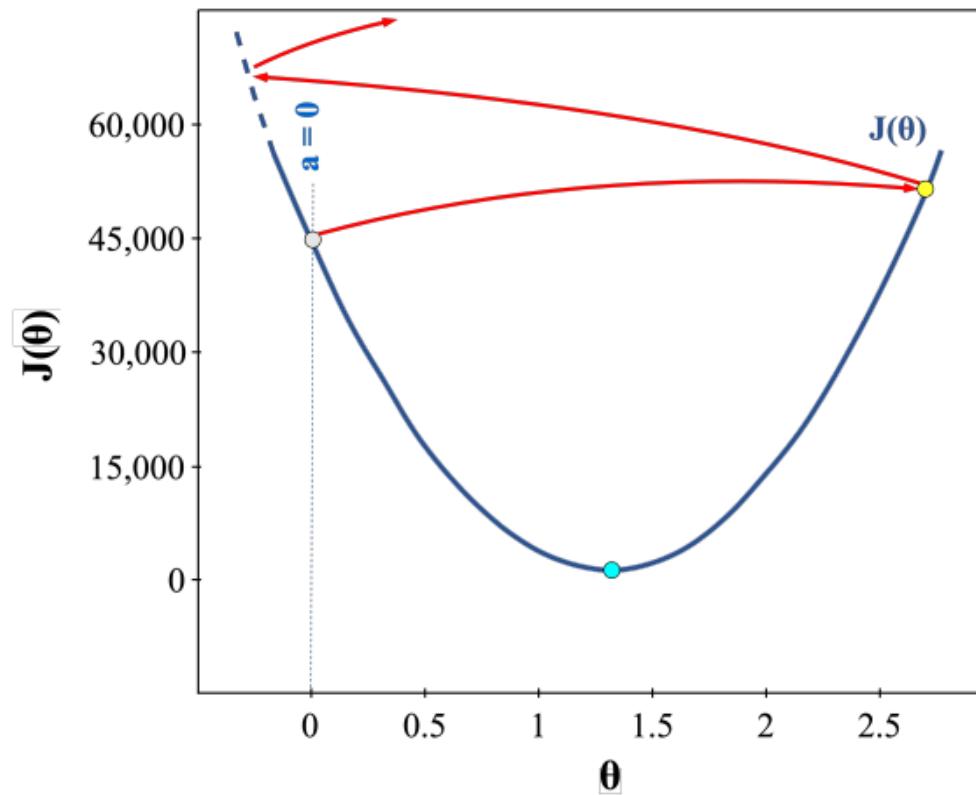
$$\text{sigmoid}'(x) = \text{sigmoid}(x) (1 - \text{sigmoid}(x))$$



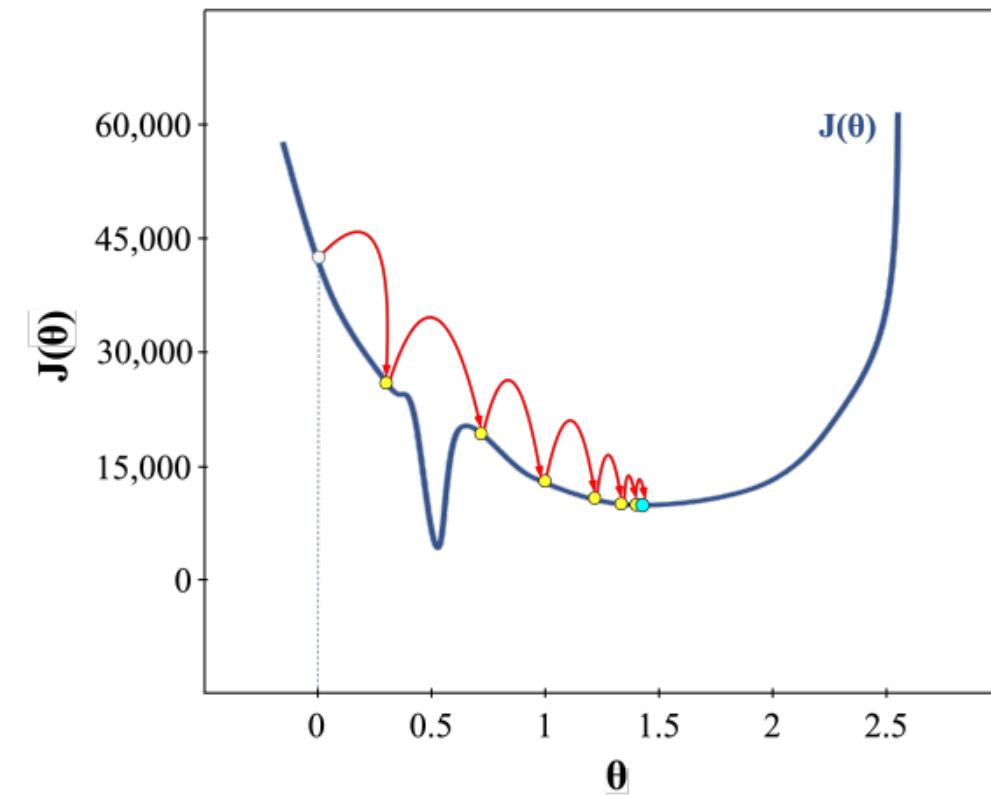
$$\tanh'(x) = 1 - \tanh^2(x)$$

Optimization Algorithms

Learning rate



(a) Gradient descent missing global minimum on a convex cost function due to a very large learning rate.



(b) Gradient Descent missing global minimum on a non-convex cost function due to a very large learning rate.

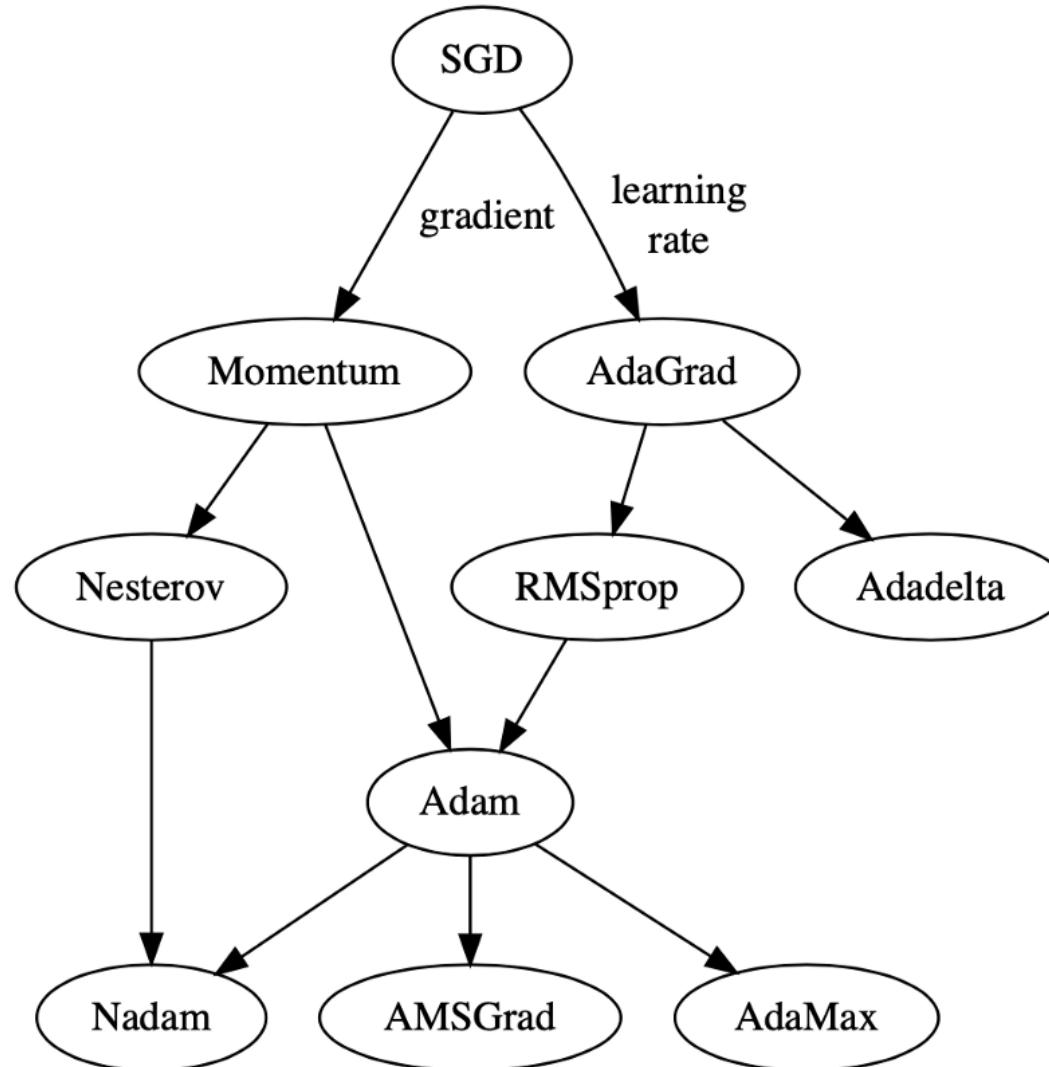
Outline

- Quick Introduction
- Stochastic Gradient Descent
- Adagrad
- RMSProp
- Momentum
- Adam
- Case studies

Optimizers

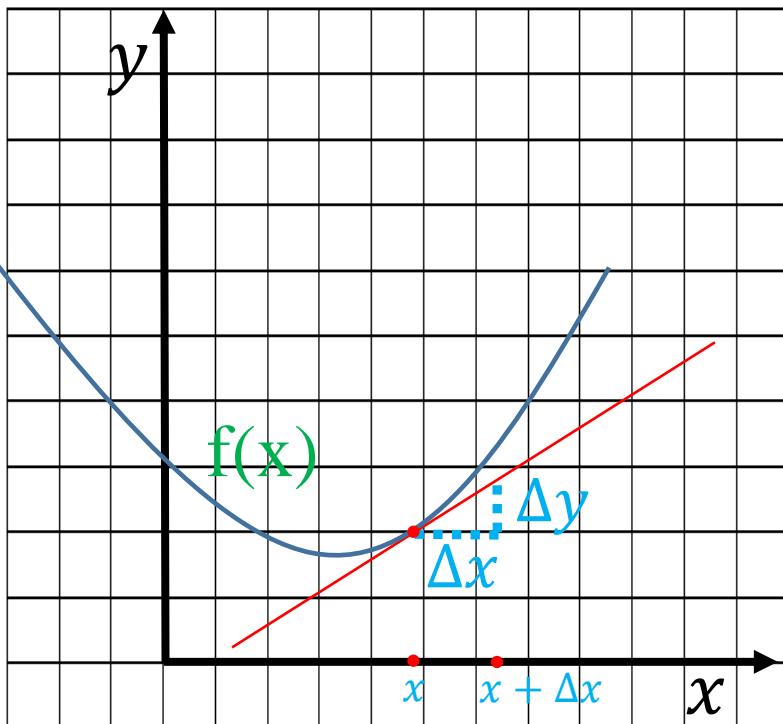
Optimizer Selection

Define a way to update parameters



Derivative/Gradient

Đạo hàm cho hàm liên tục



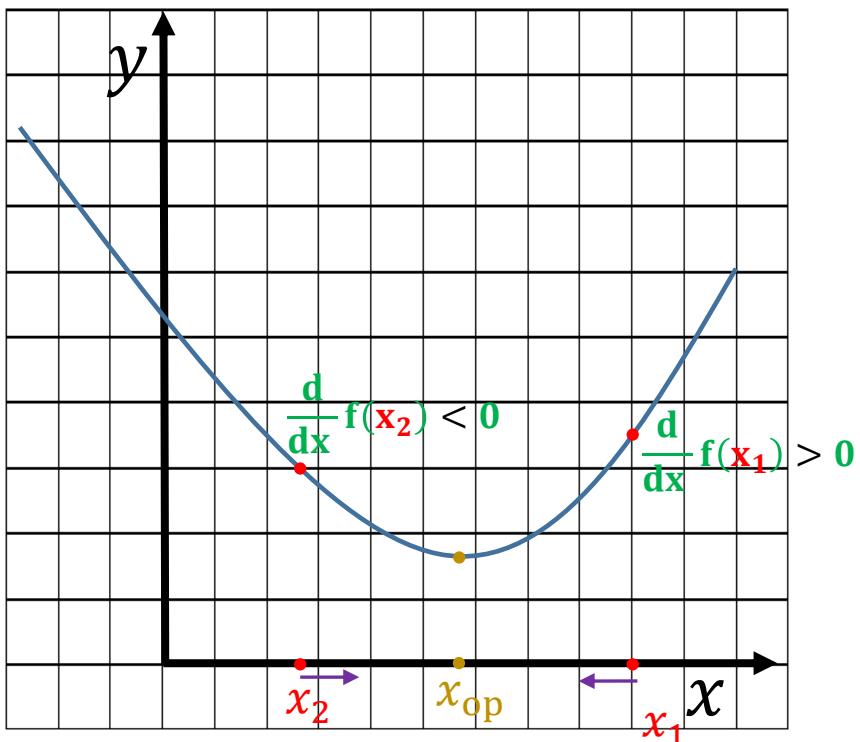
$$\text{Đạo hàm} = \frac{\text{Thay đổi theo } y}{\text{Thay đổi theo } x} = \frac{\Delta y}{\Delta x}$$

$$\frac{d}{dx} f(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Δx cần tiến về 0 để đường tiếp tuyến tiến về hàm f(x) trong vùng lân cận tại x

Gradient-based Optimization

Tìm giá trị min



Quan sát: x_{op} ở vị trí ngược hướng đạo hàm tại \mathbf{x}_1 và \mathbf{x}_2

Cách xử lý việc di chuyển ngược hướng đạo hàm cho \mathbf{x}_1 và \mathbf{x}_2 (để tìm x_{op}) khác nhau hình thành các thuật toán tối ưu hóa khác nhau

Cách cập nhật giá trị x đơn giản

$$\mathbf{x} = \mathbf{x} - \eta \frac{d}{dx} f(\mathbf{x})$$

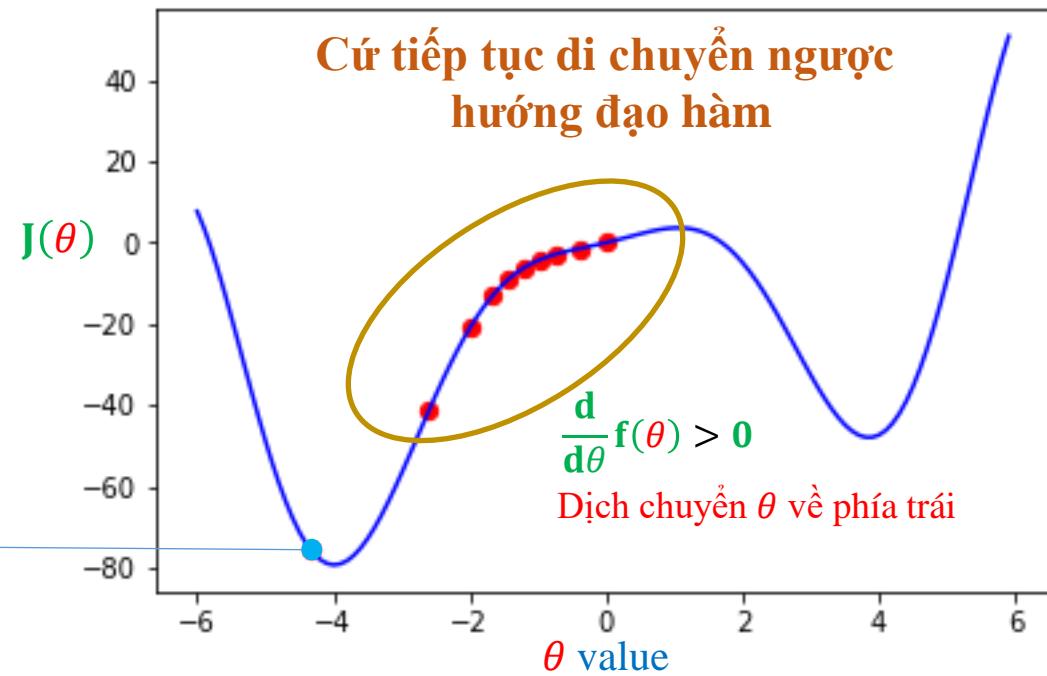
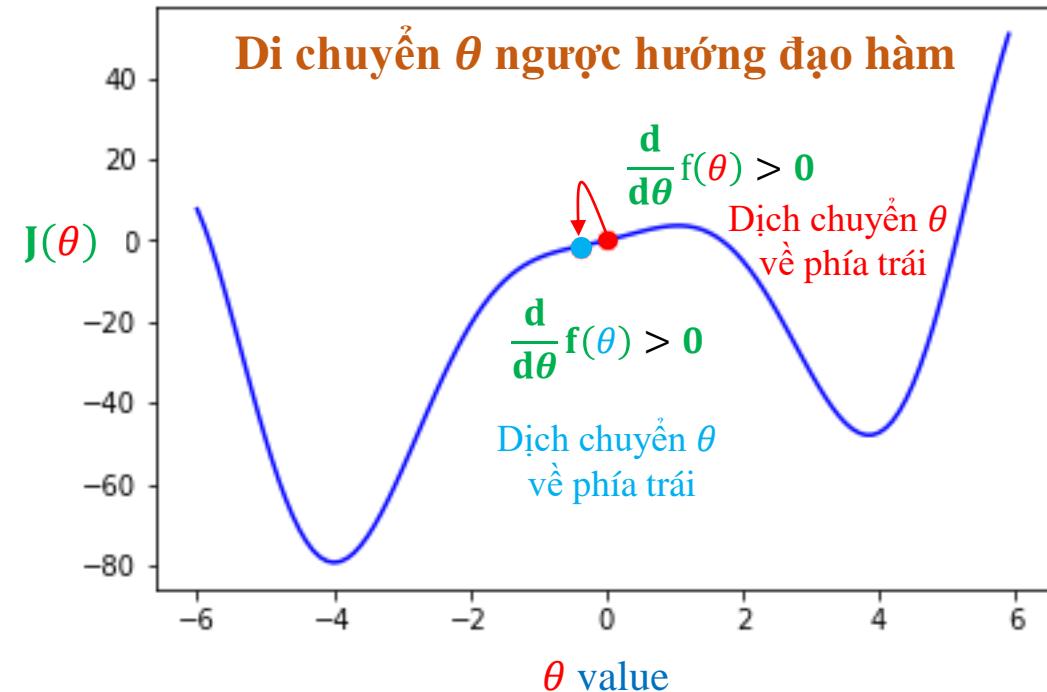
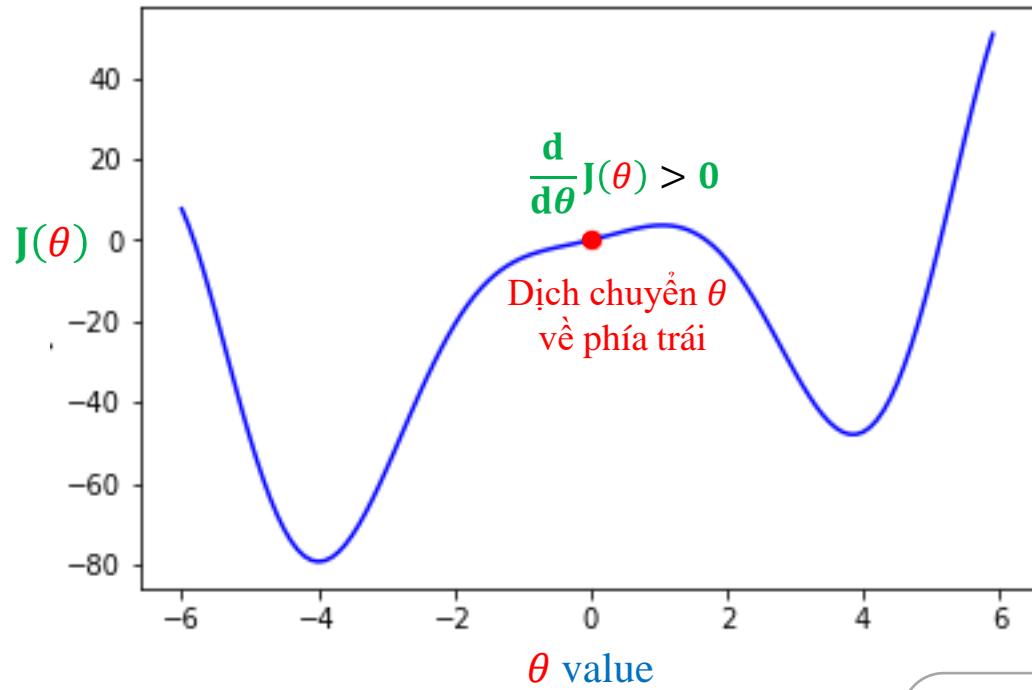
Trọng số

Đạo hàm tại x

Optimization

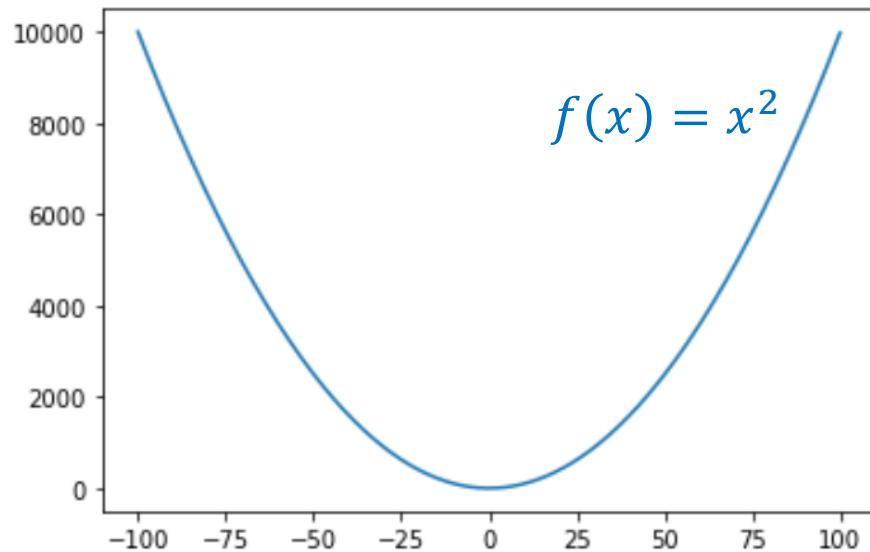
❖ A cue to optimize a function

Khởi tạo giá trị θ

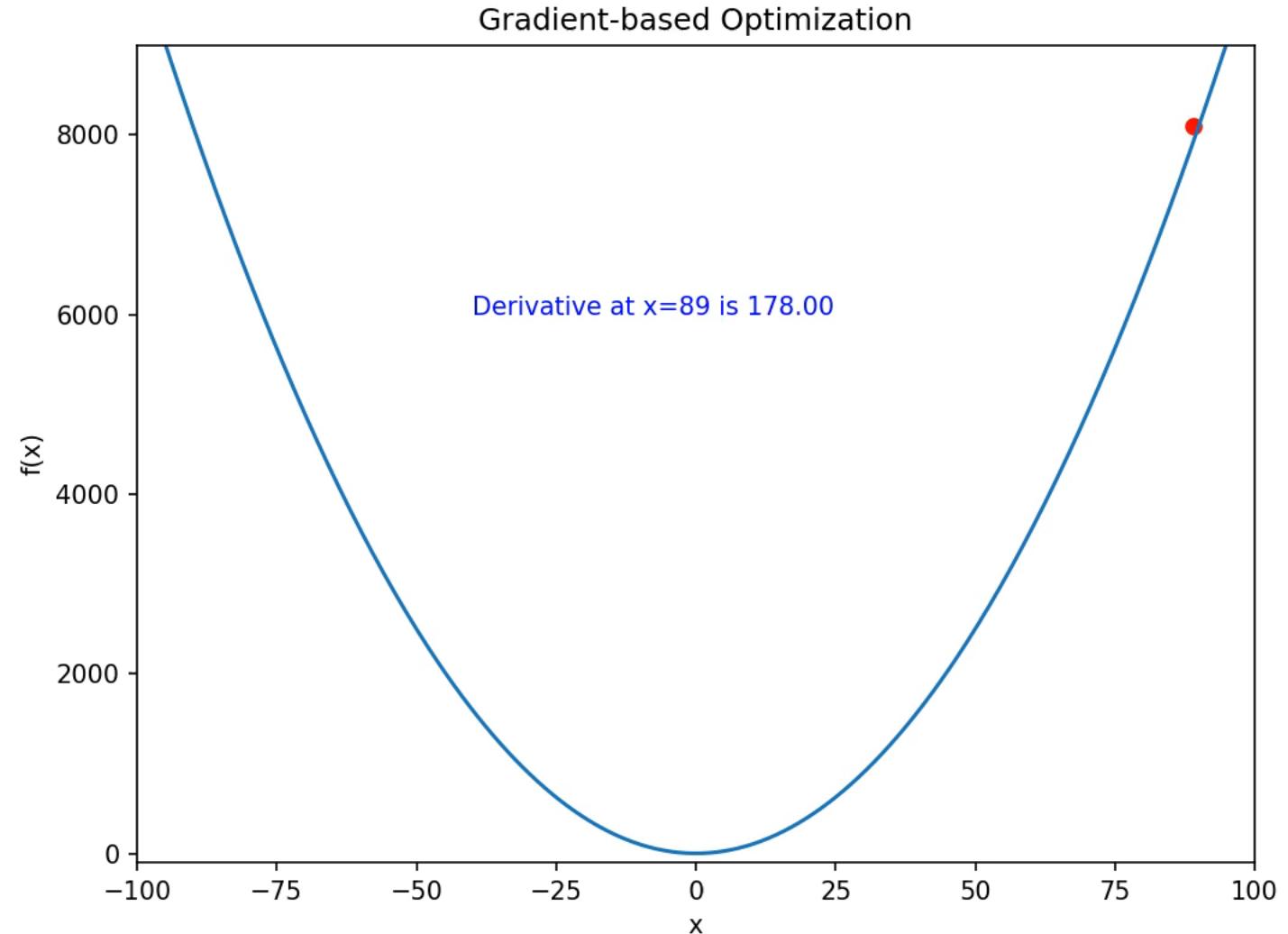


Gradient-based Optimization

❖ Square function

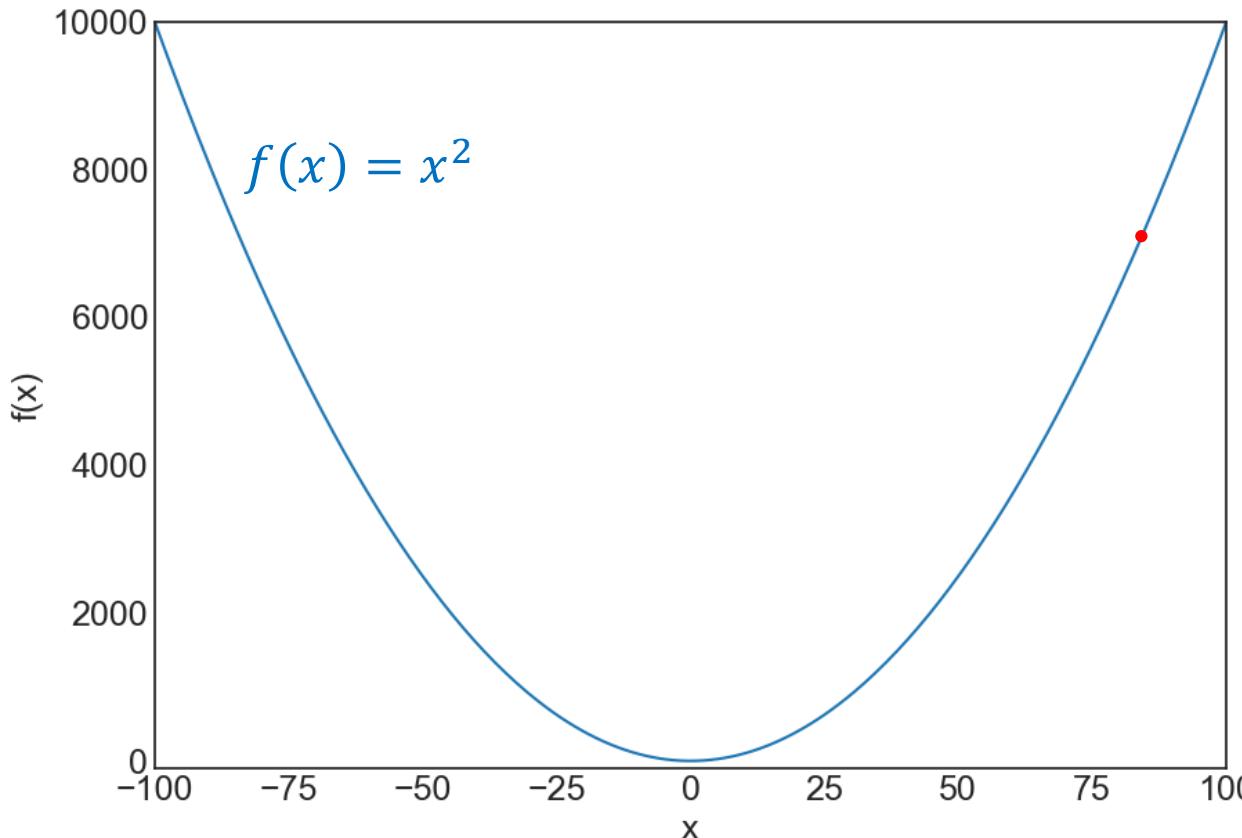


$$\begin{aligned} -100 \leq x \leq 100 \\ x \in \mathbb{N} \end{aligned}$$



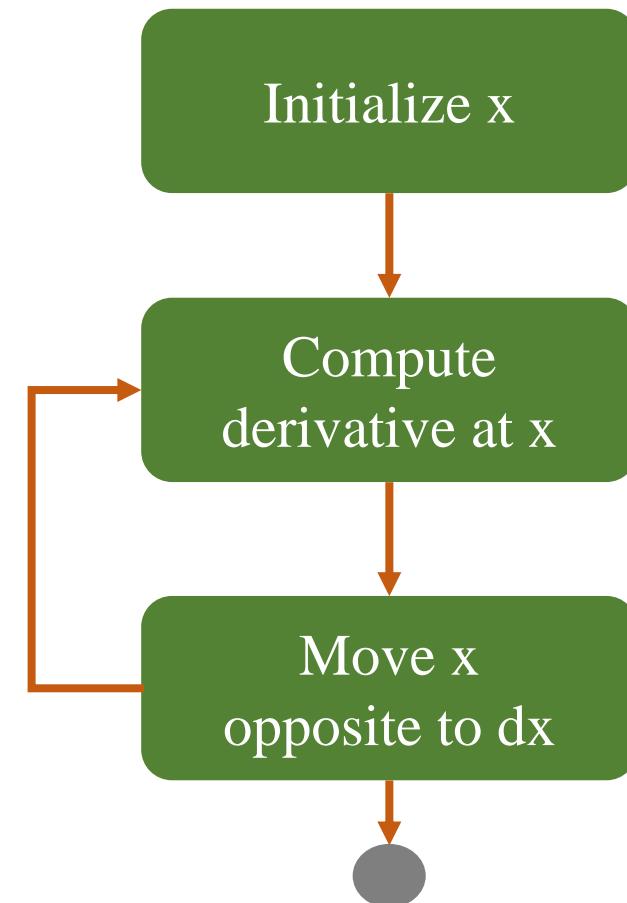
Gradient-based Optimization

❖ Square function



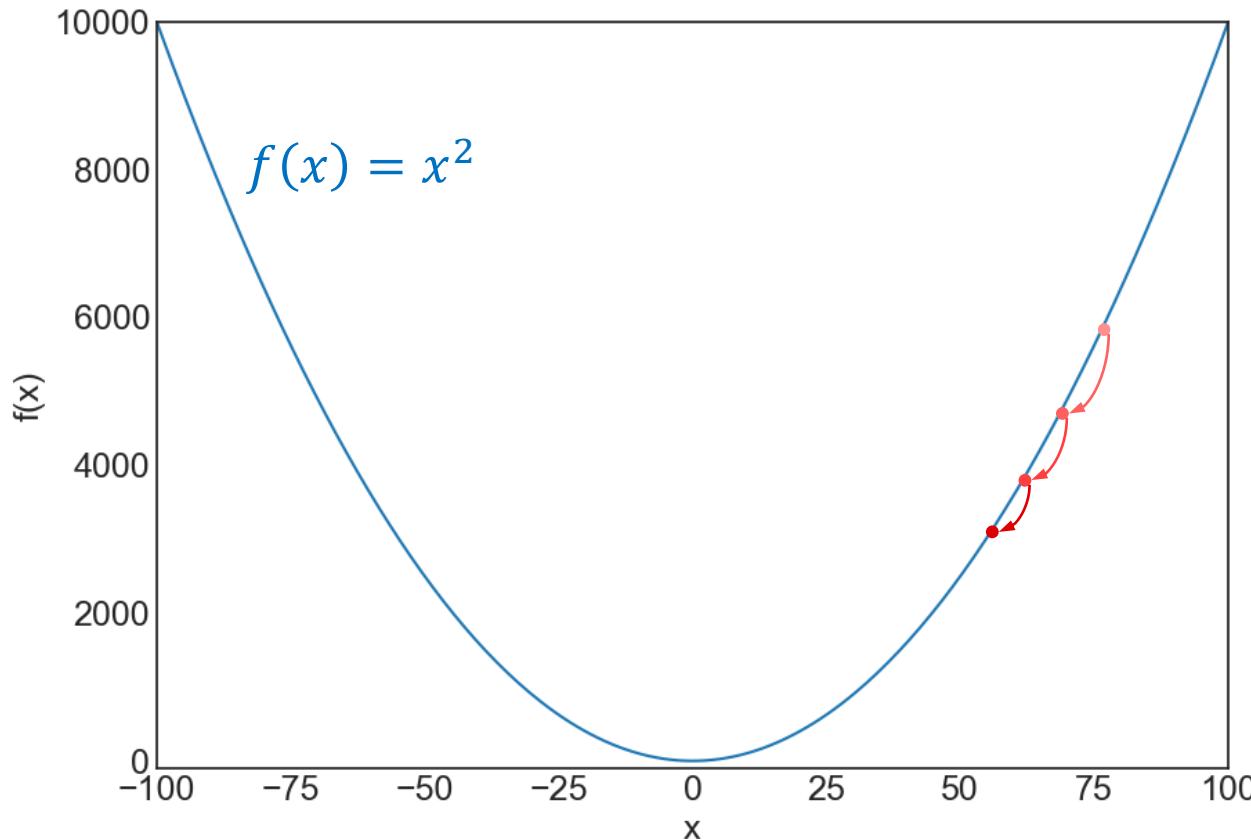
$$\begin{aligned} -100 &\leq x \leq 100 \\ x &\in \mathbb{N} \end{aligned}$$

$$x_t = x_{t-1} - \eta f'(x_{t-1})$$



Optimization

❖ Square function



$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_0 = 70.0 \quad \eta = 0.1$$

$$f'(x_0) = 140.0$$

$$x_1 = x_0 - \eta f'(x_0) = 56.0$$

$$f'(x_1) = 112.0$$

$$x_2 = x_1 - \eta f'(x_1) = 44.8$$

$$f'(x_2) = 89.6$$

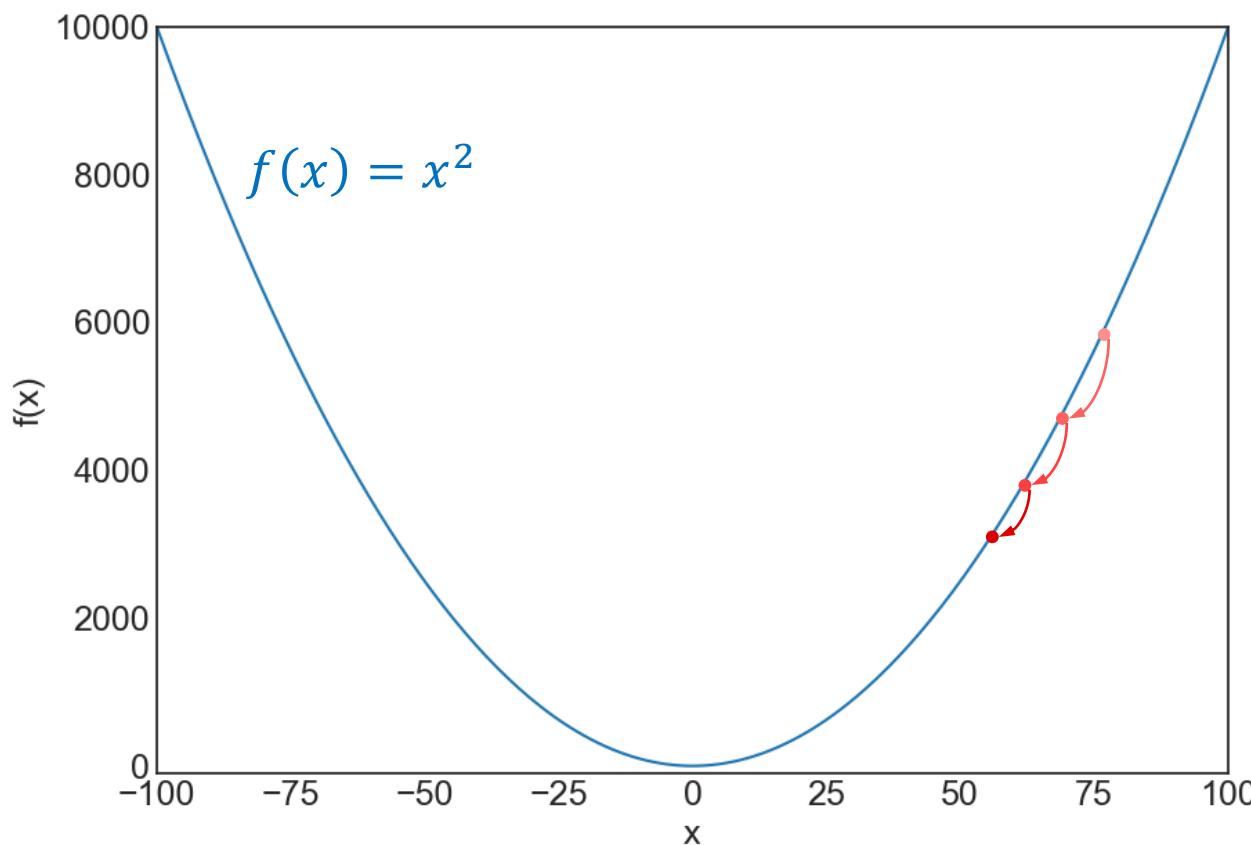
$$x_3 = x_2 - \eta f'(x_2) = 35.84$$

$$f'(x_3) = 71.68$$

$$x_4 = x_3 - \eta f'(x_3) = 28.672$$

Optimization

❖ Square function



Keep doing

$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_{10} = 6.012 \quad \eta = 0.1$$

$$f'(x_{10}) = 12.02$$

$$x_{11} = x_{10} - \eta f'(x_{10}) = 4.81$$

$$f'(x_{11}) = 9.62$$

$$x_{12} = x_{11} - \eta f'(x_{11}) = 3.84$$

$$f'(x_{12}) = 7.69$$

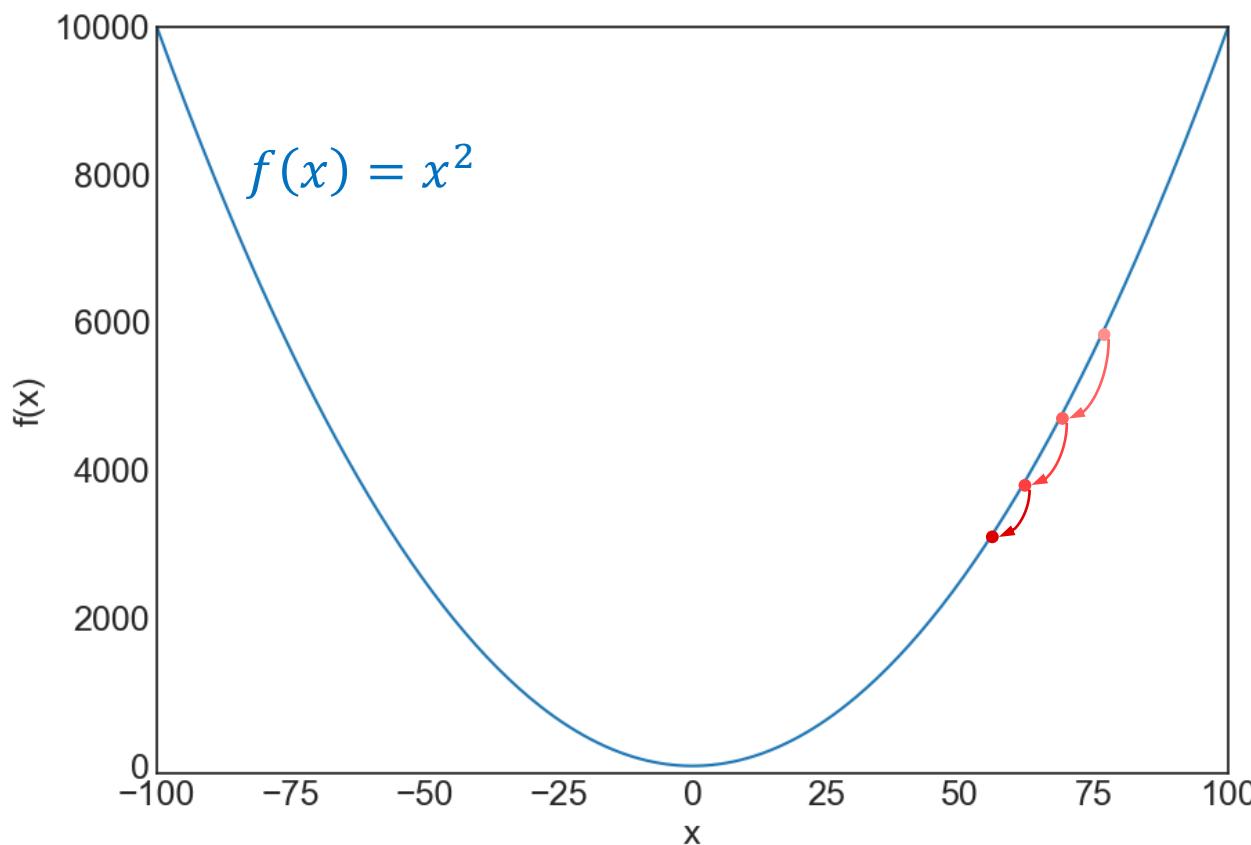
$$x_{13} = x_{12} - \eta f'(x_{12}) = 3.078$$

$$f'(x_{13}) = 6.15$$

$$x_{14} = x_{13} - \eta f'(x_{13}) = 2.46$$

Optimization

❖ Square function



Keep doing

$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_{30} = 0.069 \quad \eta = 0.1$$

$$f'(x_{30}) = 0.138$$

$$x_{31} = x_{30} - \eta f'(x_{30}) = 0.055$$

$$f'(x_{31}) = 0.11$$

$$x_{32} = x_{31} - \eta f'(x_{31}) = 0.044$$

$$f'(x_{32}) = 0.88$$

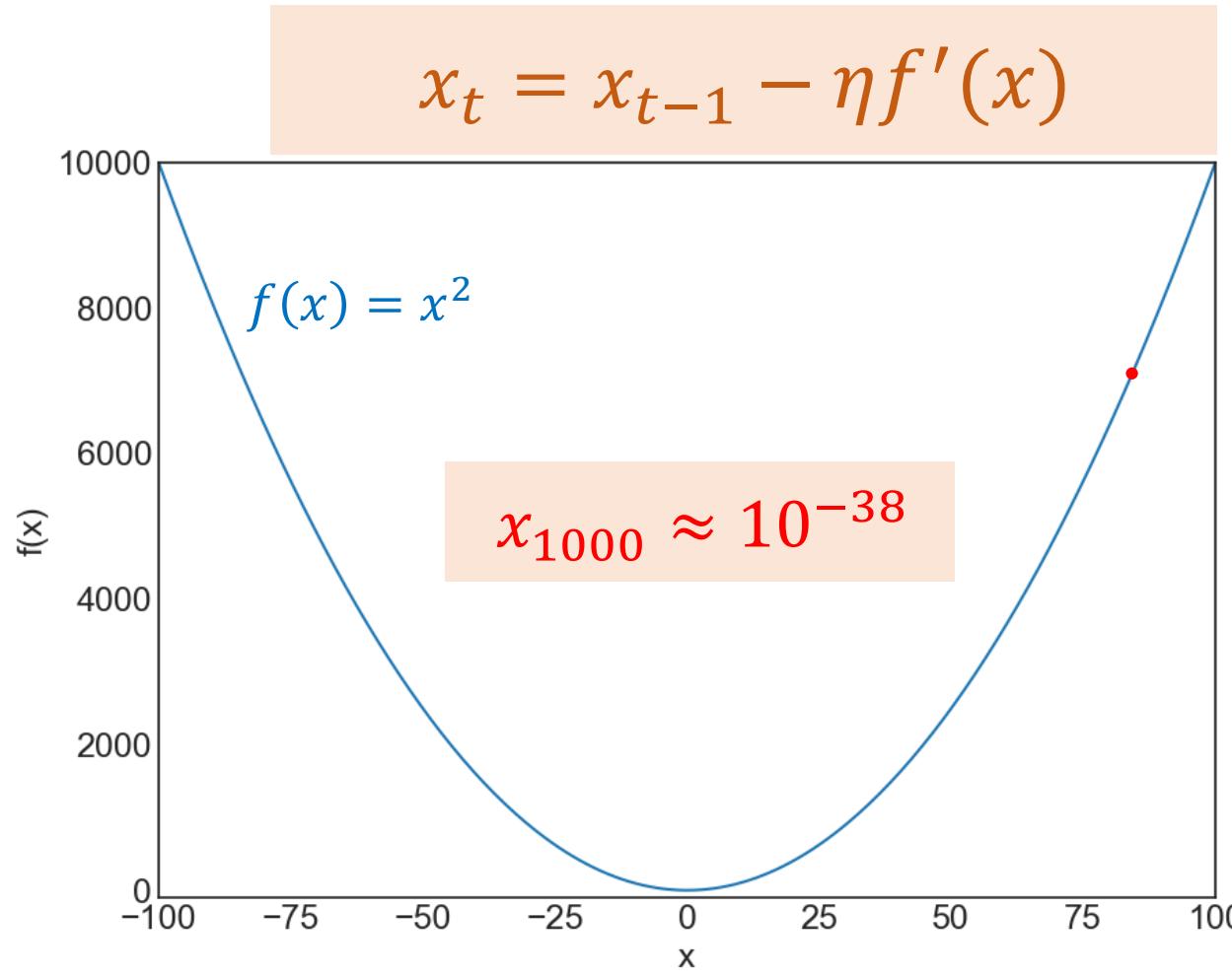
$$x_{33} = x_{32} - \eta f'(x_{32}) = 0.035$$

$$f'(x_{34}) = 0.071$$

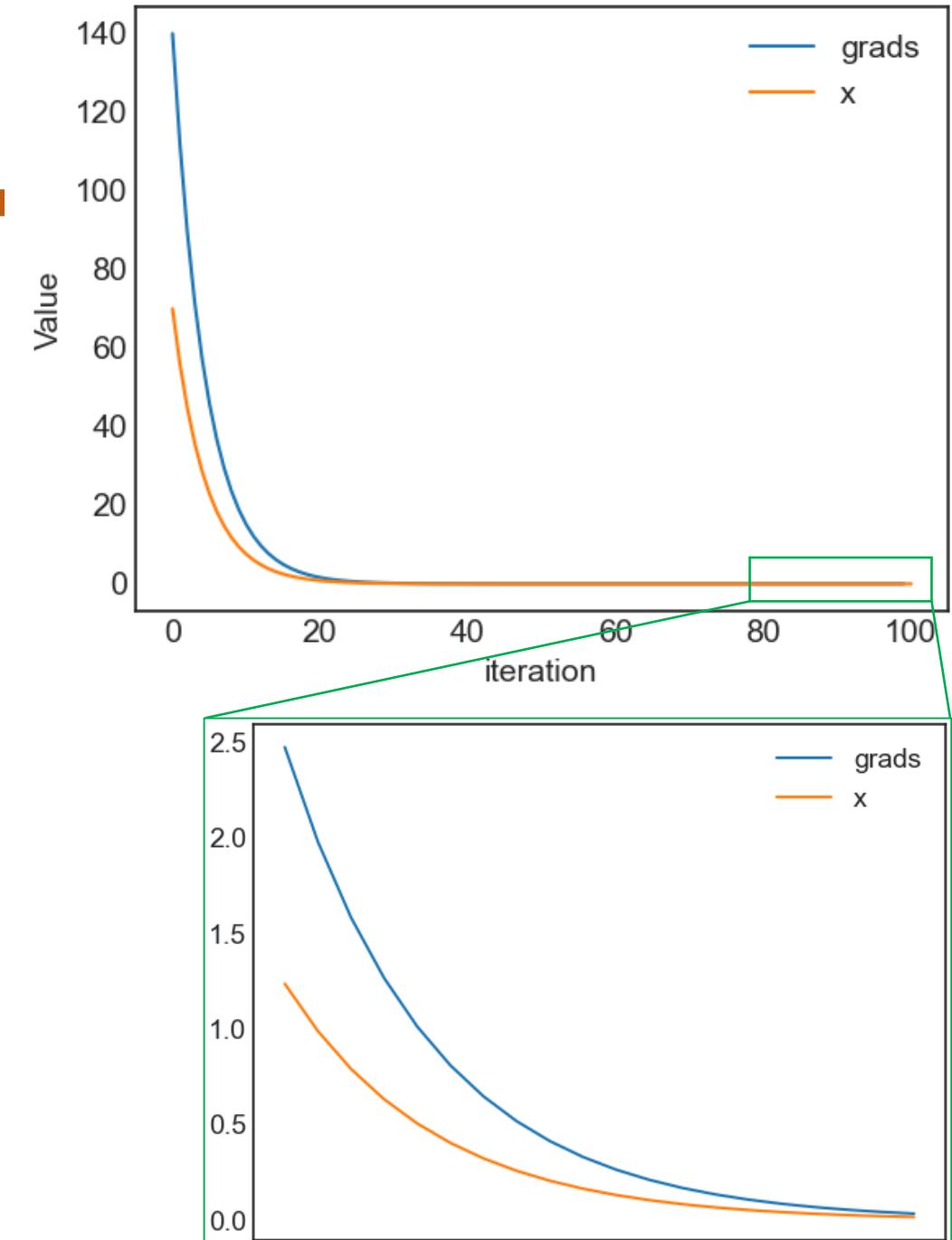
$$x_{34} = x_{33} - \eta f'(x_{33}) = 0.028$$

Optimization

❖ Square function



Optimized successfully!



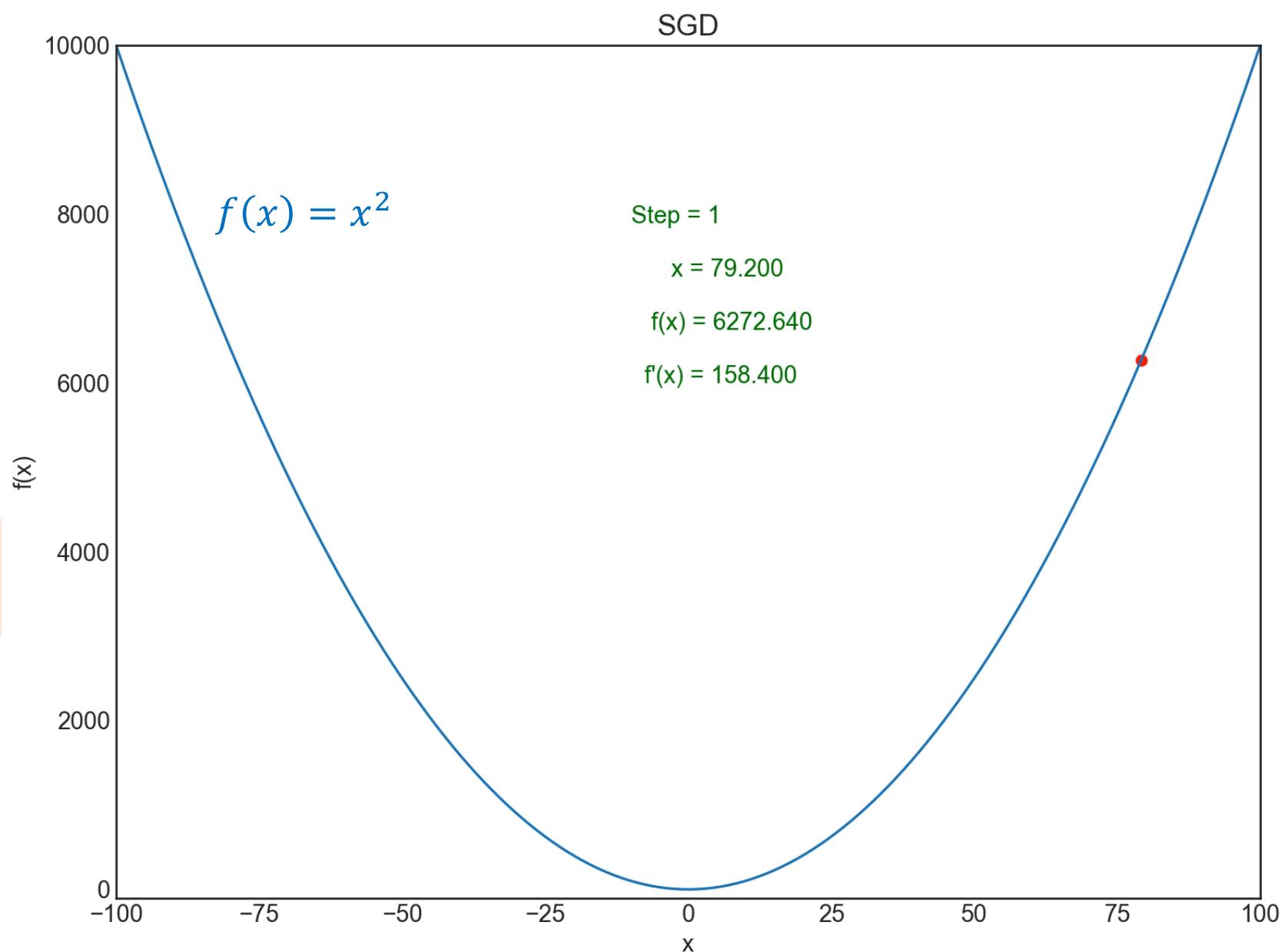
Optimization

❖ Square function

$$x_0 = 99.0$$

$$\eta = 0.1$$

$$x_t = x_{t-1} - \eta f'(x)$$



Optimization

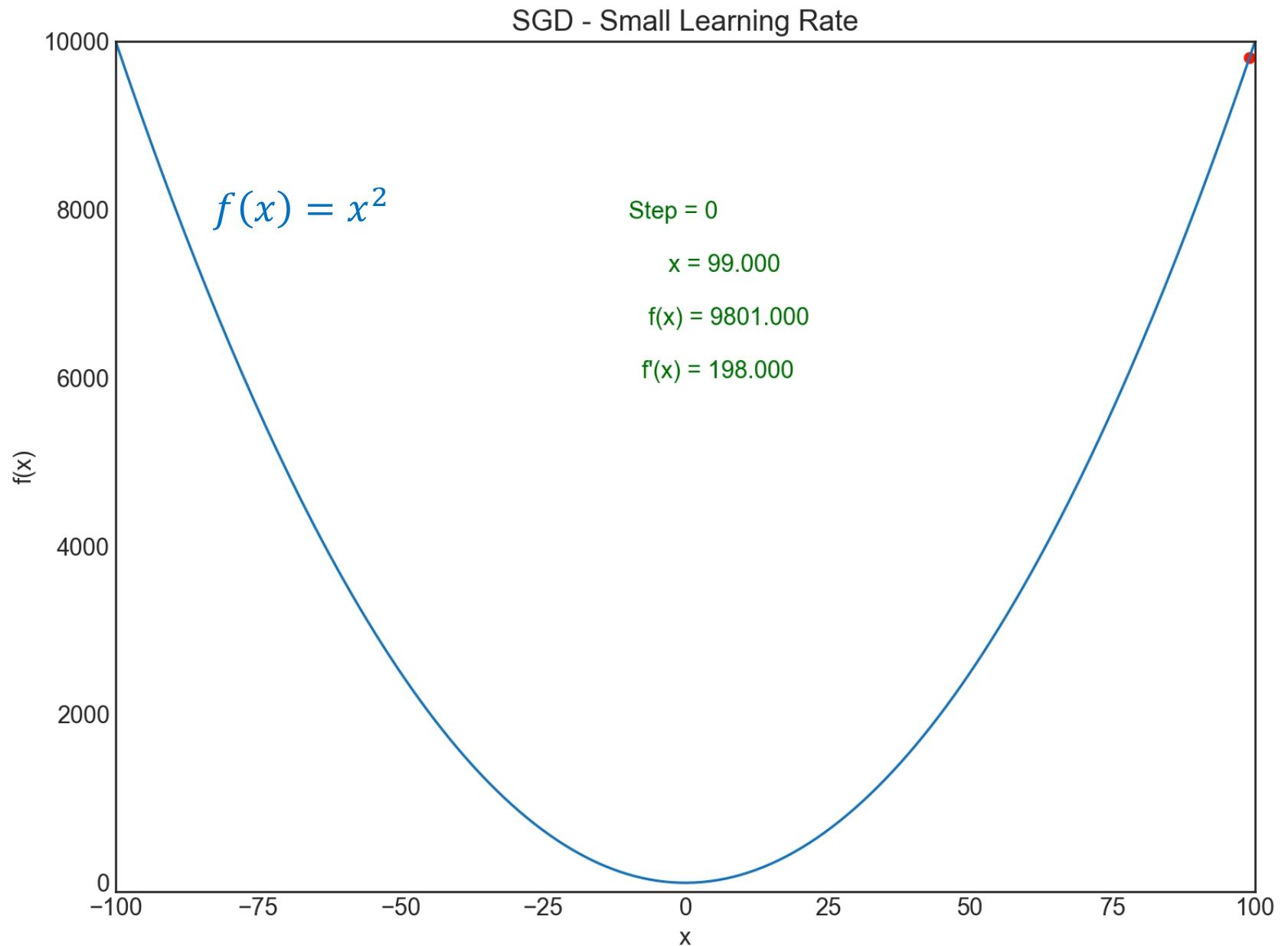
❖ Square function

Discussion

$$x_0 = 99.0$$

$$\eta = 0.001$$

$$x_t = x_{t-1} - \eta f'(x)$$



Optimization

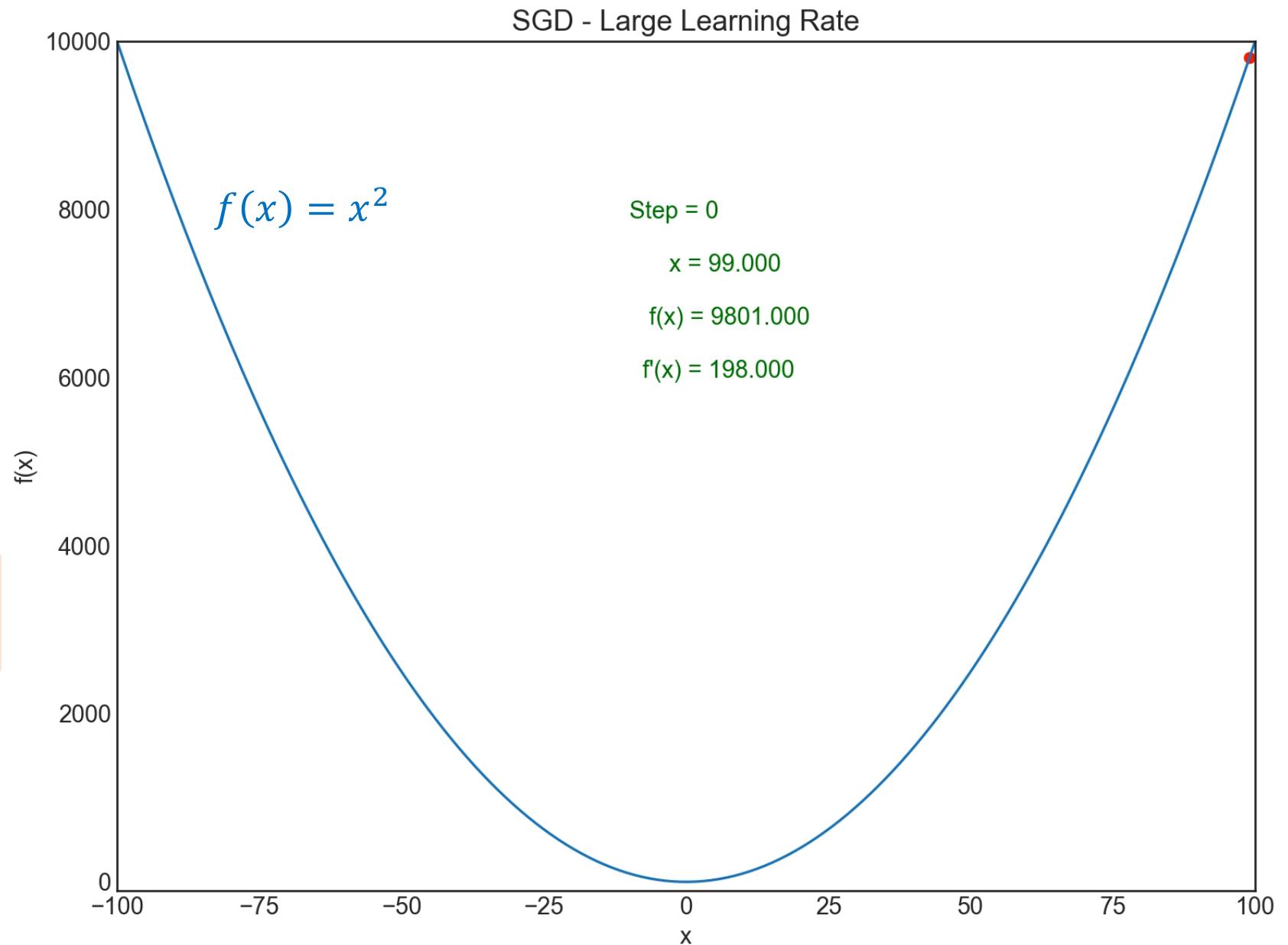
❖ Square function

Discussion

$$x_0 = 99.0$$

$$\eta = 0.8$$

$$x_t = x_{t-1} - \eta f'(x)$$



Optimization

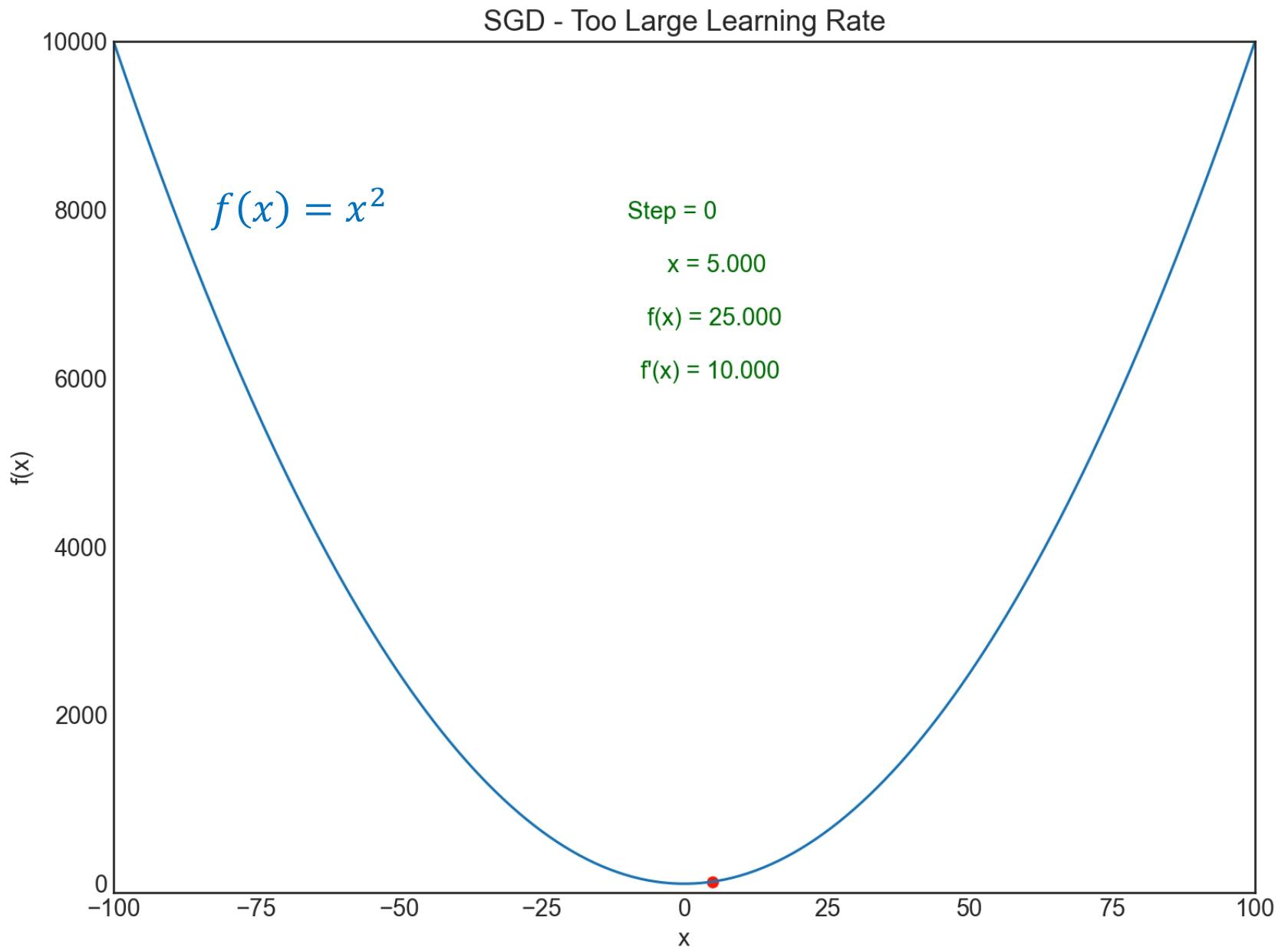
❖ Square function

Discussion

$$x_0 = 99.0$$

$$\eta = 1.1$$

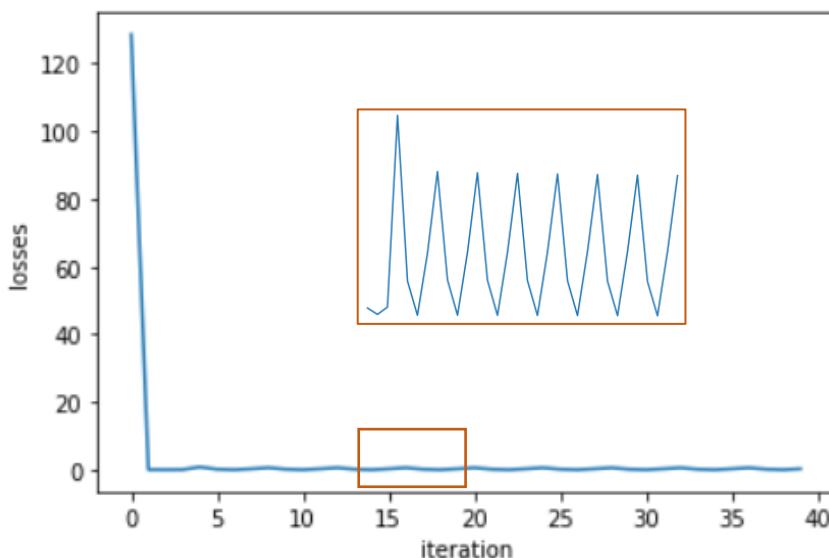
$$x_t = x_{t-1} - \eta f'(x)$$



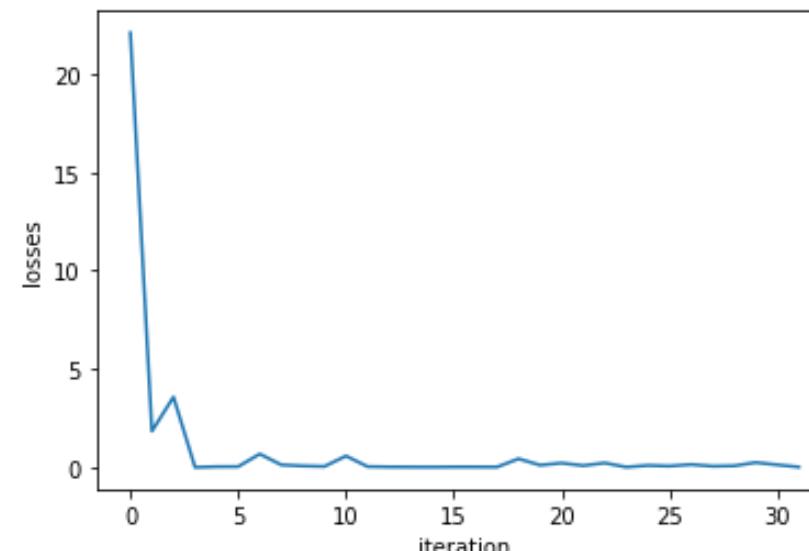
Optimization Algorithms

Stochastic gradient descent

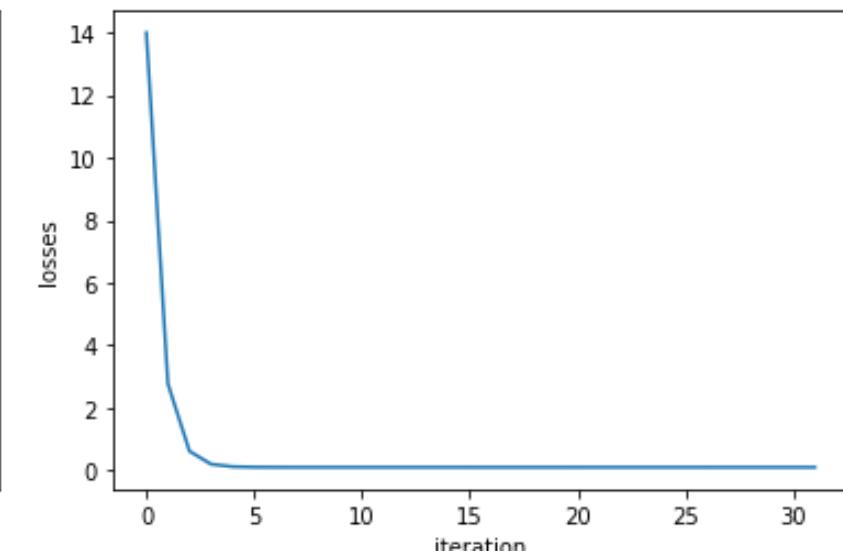
$$\theta_t = \theta_{t-1} - \eta \nabla_{\theta} L$$



1-sample



m-sample



N-sample

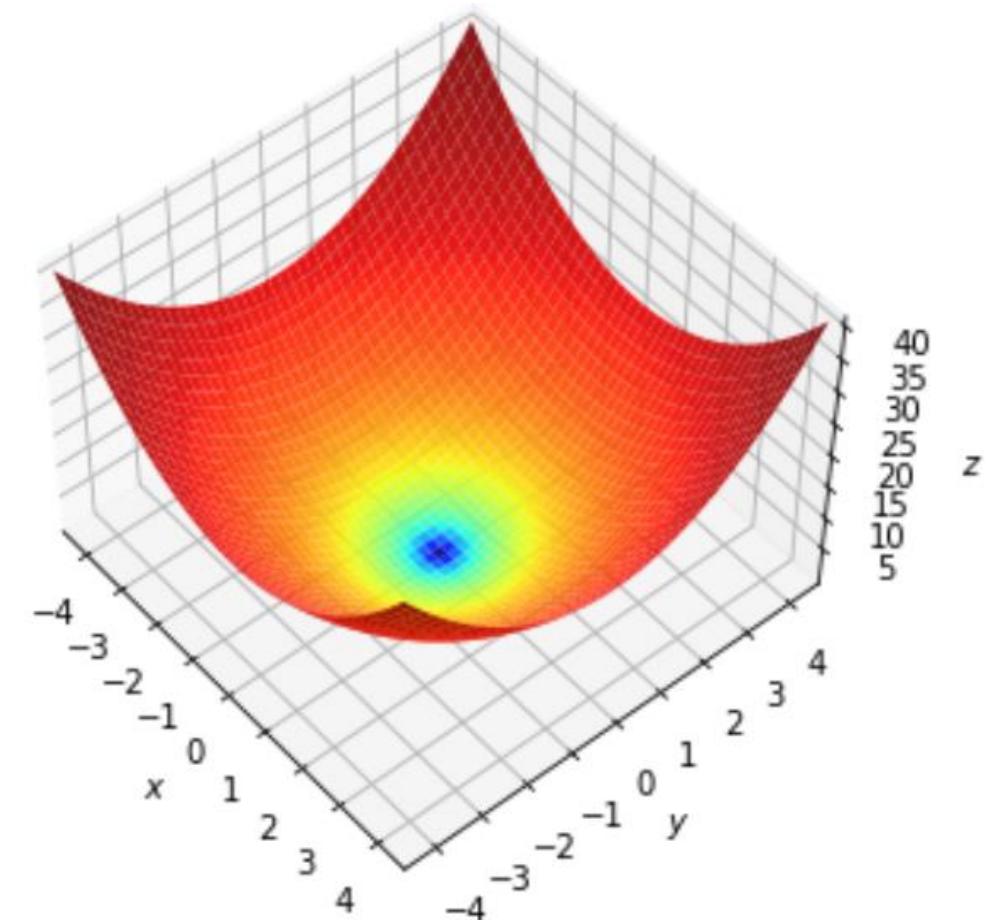
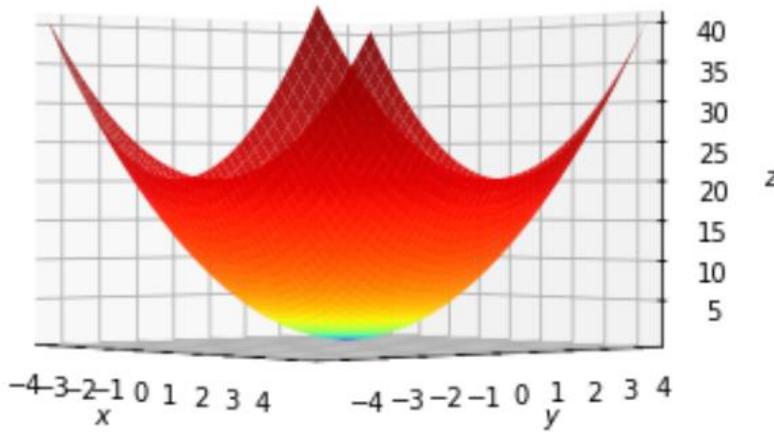
Optimization

❖ Optimization: 2D function

$$f(x, y) = x^2 + y^2$$

$$-100 \leq x, y \leq 100$$

$$x, y \in \mathbb{N}$$



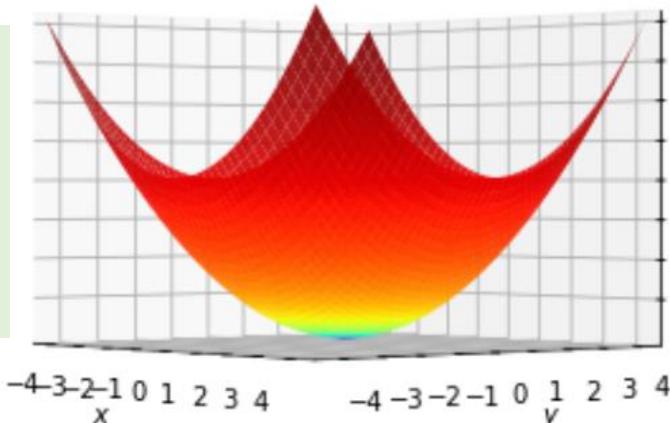
Optimization

❖ Optimization: 2D function

$$f(x, y) = x^2 + y^2$$

$$-100 \leq x, y \leq 100$$

$$x, y \in \mathbb{N}$$



$$x = x - \eta \frac{\partial f(x, y)}{\partial x}$$

$$y = y - \eta \frac{\partial f(x, y)}{\partial y}$$

$$x_0 = 6.0 \quad y_0 = 9.0 \quad \eta = 0.1$$

$$\frac{\partial f(x_0, y_0)}{\partial x} = 12 \quad \frac{\partial f(x_0, y_0)}{\partial y} = 18$$

$$x_1 = 4.8$$

$$y_1 = 7.2$$

$$\frac{\partial f(x_1, y_1)}{\partial x} = 9.6 \quad \frac{\partial f(x_1, y_1)}{\partial y} = 14.4$$

$$x_2 = 3.84$$

$$y_2 = 5.75$$

$$\frac{\partial f(x_2, y_2)}{\partial x} = 7.68 \quad \frac{\partial f(x_2, y_2)}{\partial y} = 11.51$$

$$x_3 = 3.07$$

$$y_3 = 4.608$$

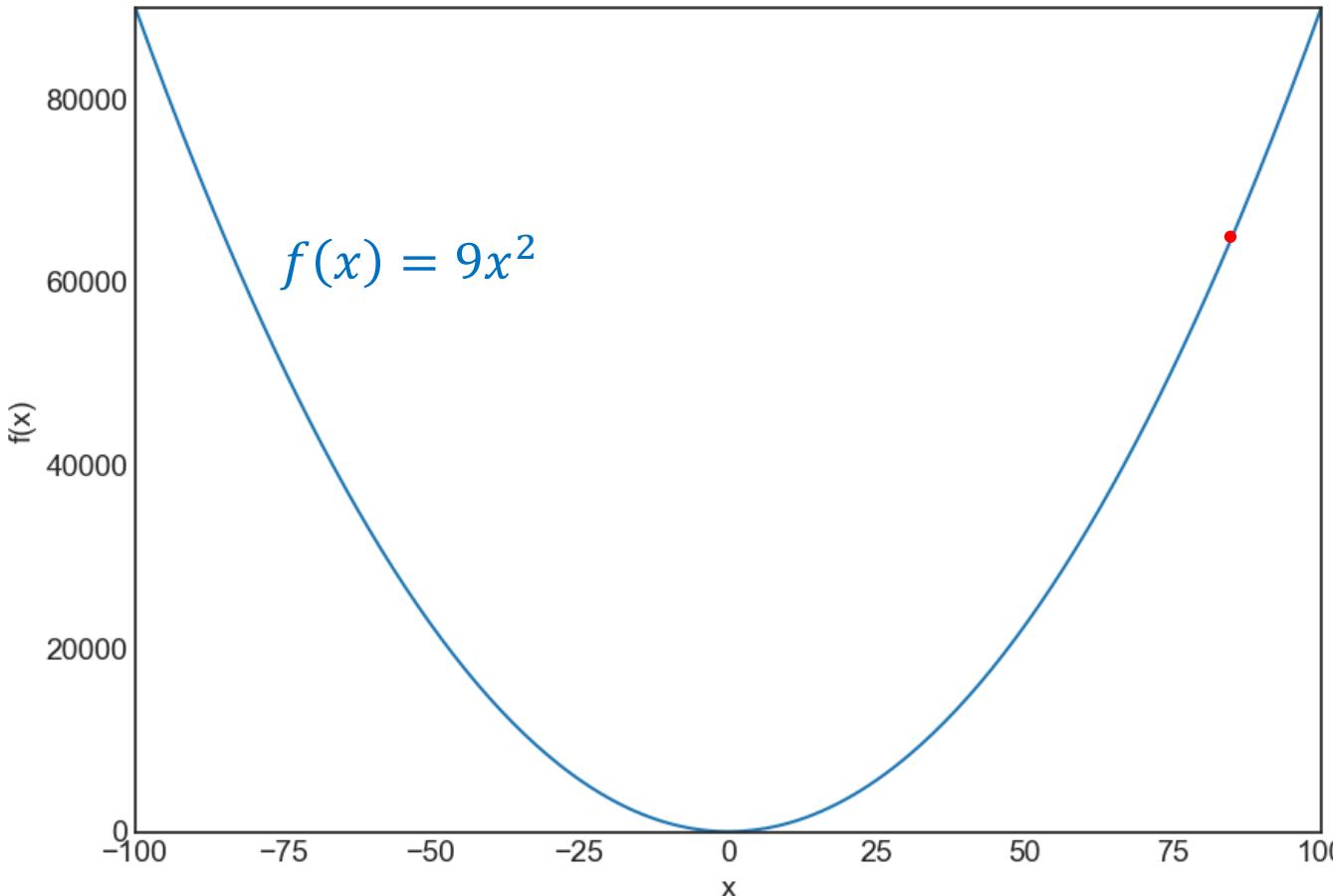
$$\frac{\partial f(x_3, y_3)}{\partial x} = 6.14 \quad \frac{\partial f(x_3, y_3)}{\partial y} = 9.21$$

$$x_4 = 2.45$$

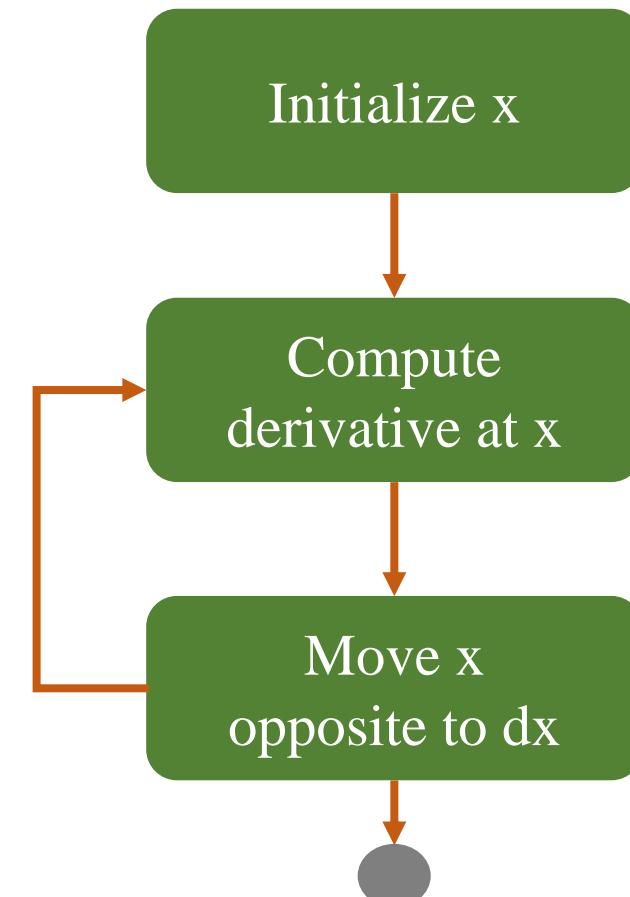
$$y_4 = 3.68$$

Gradient-based Optimization

❖ Another Square function

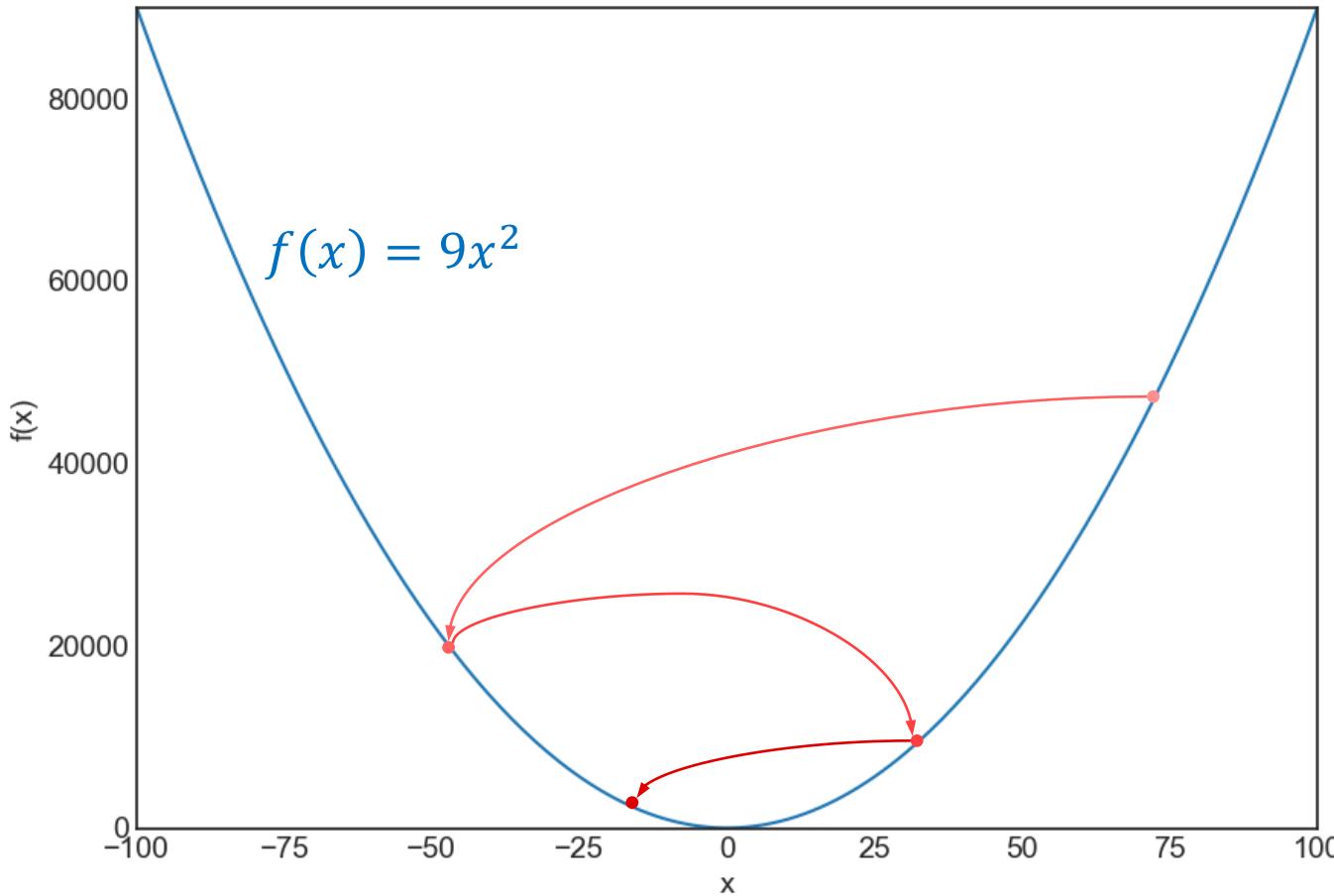


$$x_t = x_{t-1} - \eta f'(x)$$



Optimization

❖ Another Square function



$$-100 \leq x \leq 100 \\ x \in \mathbb{N}$$

$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_0 = 70.0 \quad \eta = 0.1$$

$$f'(x_0) = 1260.0$$

$$x_1 = x_0 - \eta f'(x_0) = -56.0$$

$$f'(x_1) = -1008.0$$

$$x_2 = x_1 - \eta f'(x_1) = 44.8$$

$$f'(x_2) = 806.4$$

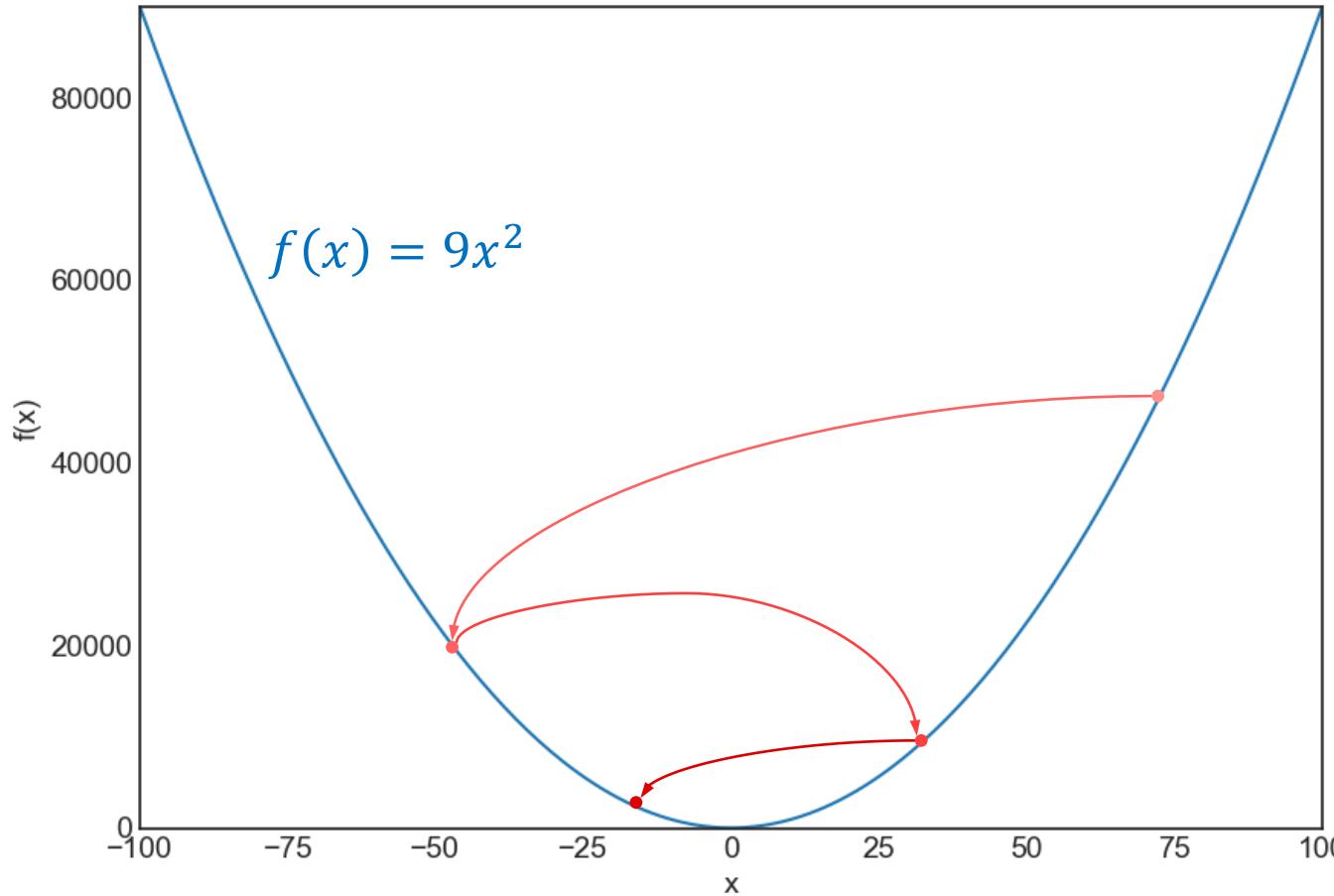
$$x_3 = x_2 - \eta f'(x_2) = -35.84$$

$$f'(x_3) = -645.12$$

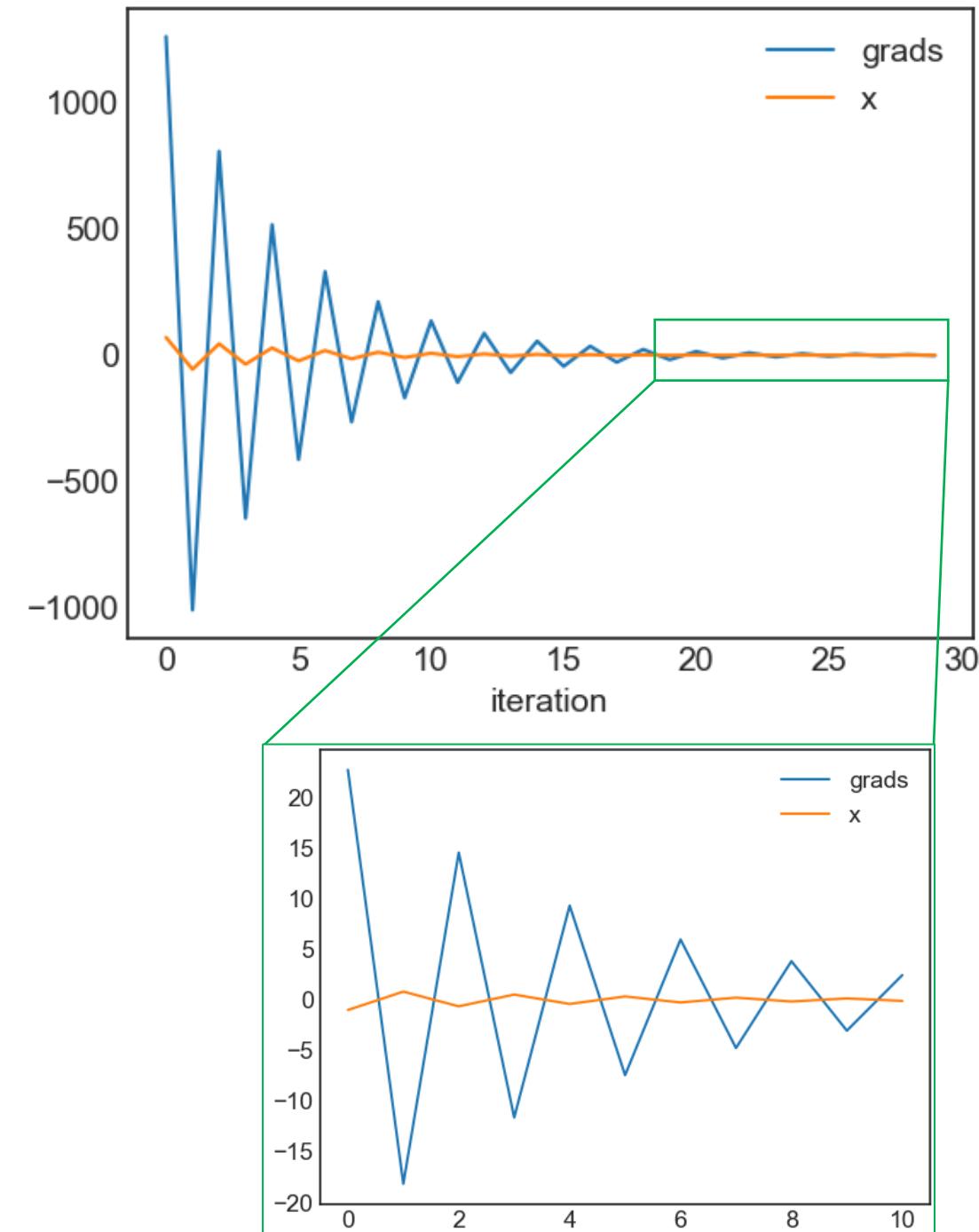
$$x_4 = x_3 - \eta f'(x_3) = 28.672$$

Optimization

❖ Another Square function



$$-100 \leq x \leq 100$$
$$x \in \mathbb{N}$$



Optimization

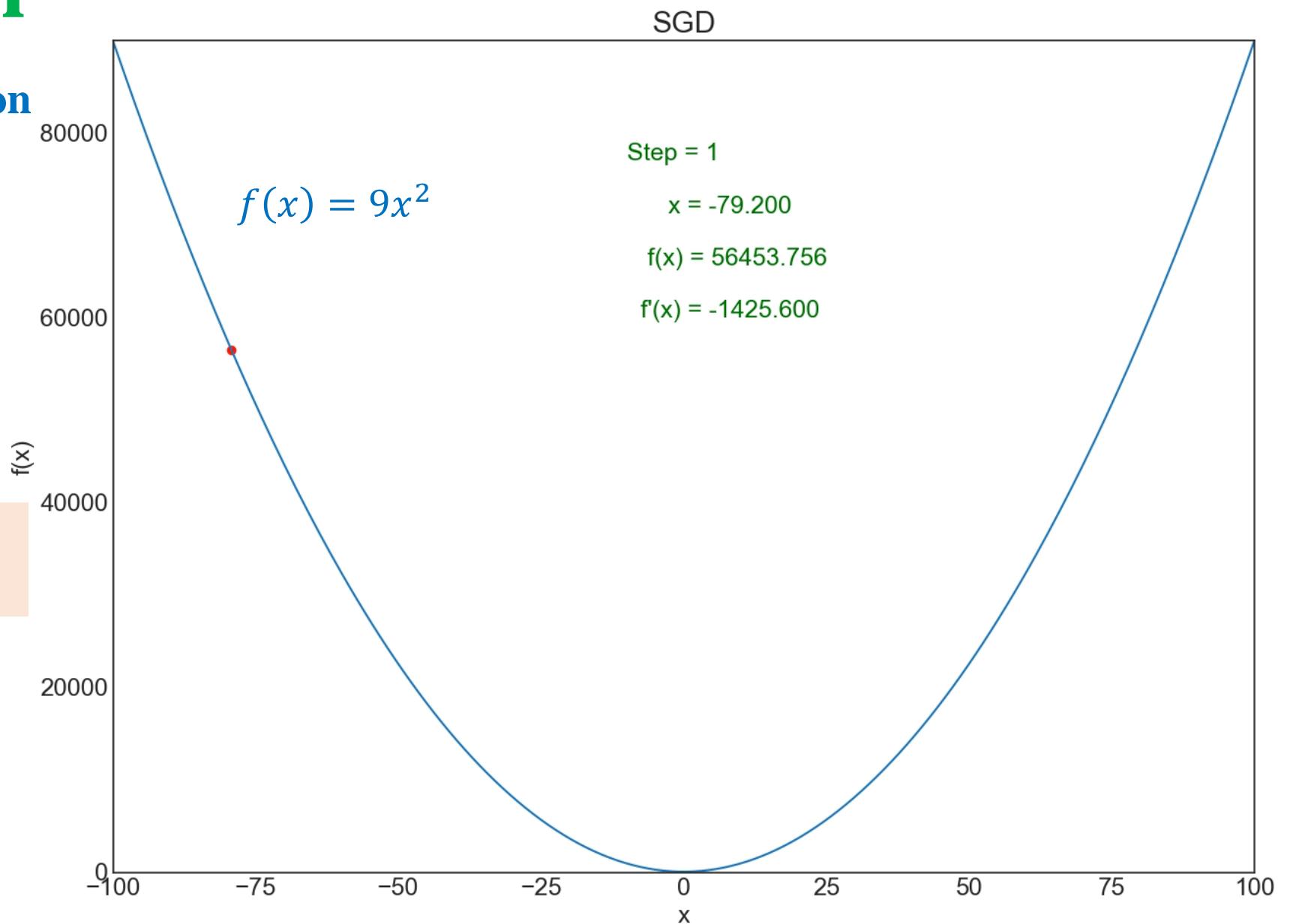
❖ Another Square function

$$x_0 = 99.0$$

$$\eta = 0.1$$

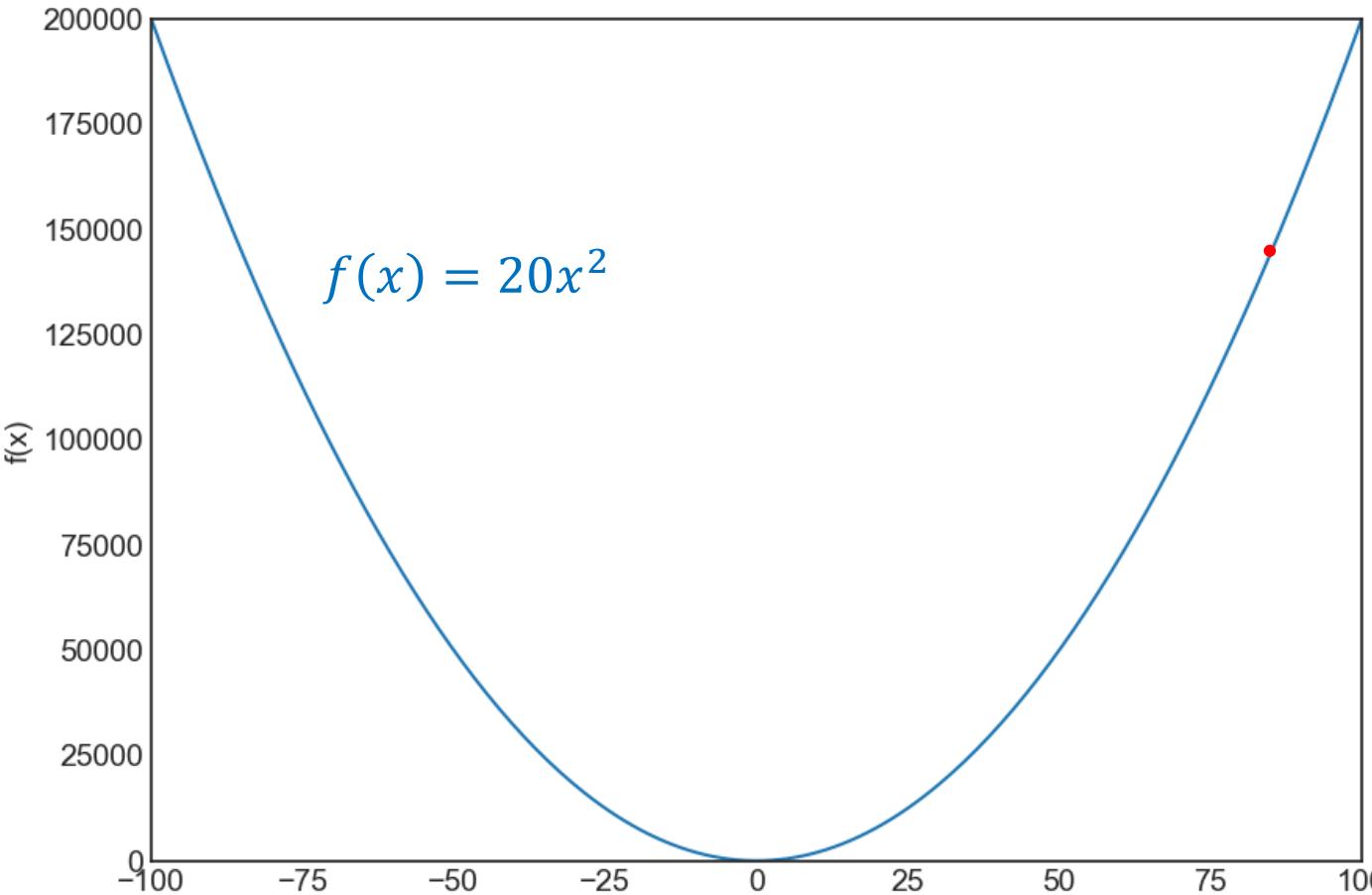
$$x_t = x_{t-1} - \eta f'(x)$$

Observation?

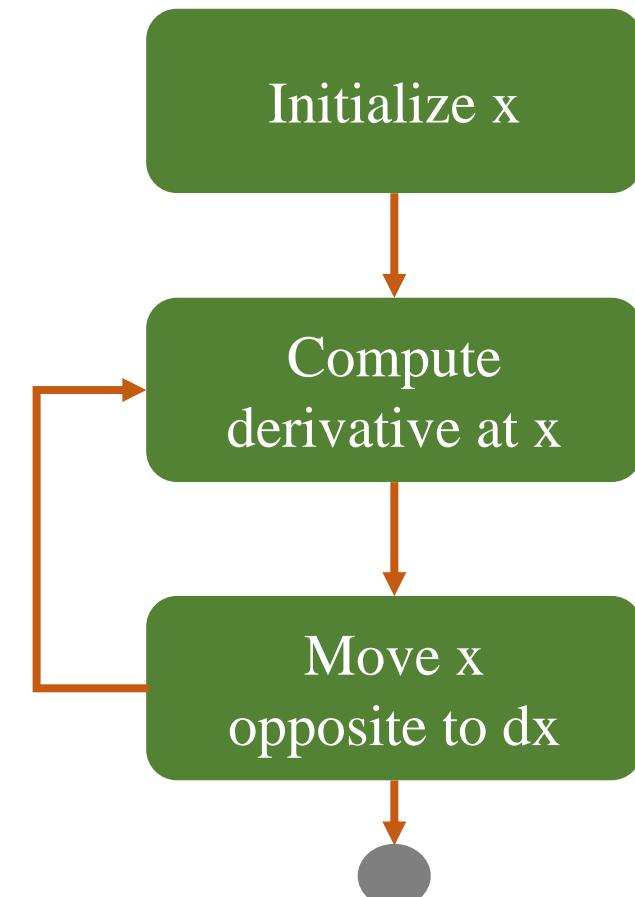


Gradient-based Optimization

❖ Square function

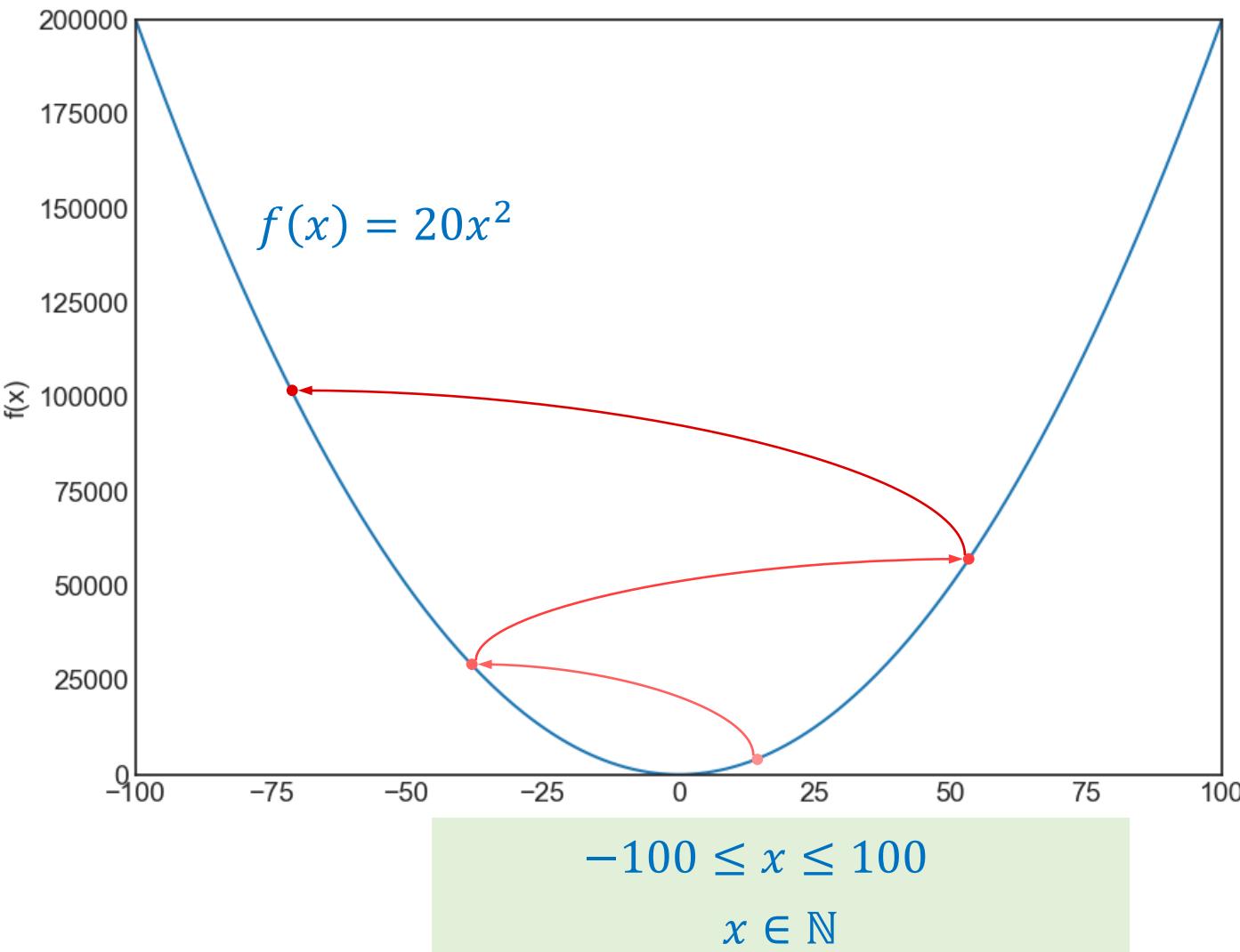


$$x_t = x_{t-1} - \eta f'(x)$$



Optimization

❖ Square function



$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_0 = 2.0 \quad \eta = 0.1$$

$$f'(x_0) = 80.0$$

$$x_1 = x_0 - \eta f'(x_0) = -6.0$$

$$f'(x_1) = -240.0$$

$$x_2 = x_1 - \eta f'(x_1) = 18.0$$

$$f'(x_2) = 720.0$$

$$x_3 = x_2 - \eta f'(x_2) = -54.0$$

$$f'(x_3) = -2160.0$$

$$x_4 = x_3 - \eta f'(x_3) = 162.0$$

Optimization

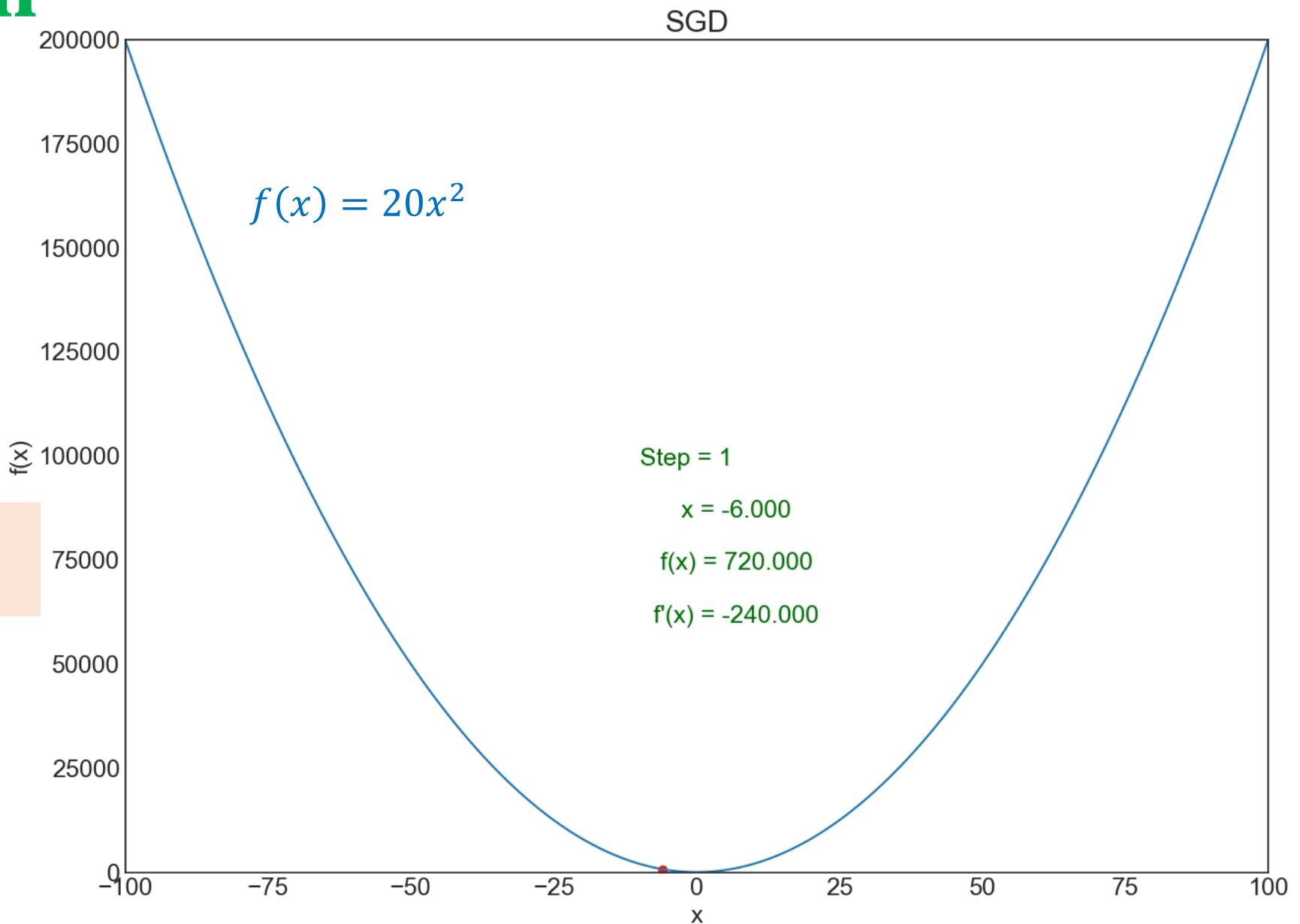
❖ Square function

$$x_0 = 2.0$$

$$\eta = 0.1$$

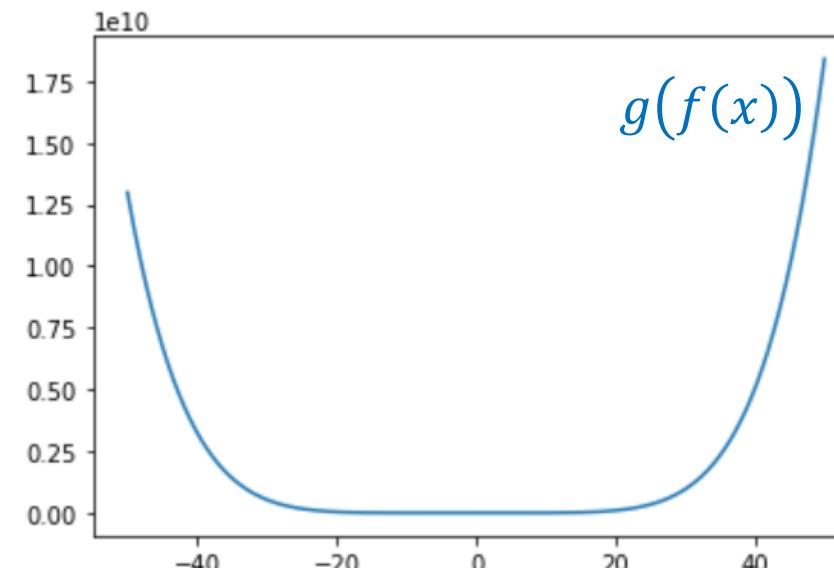
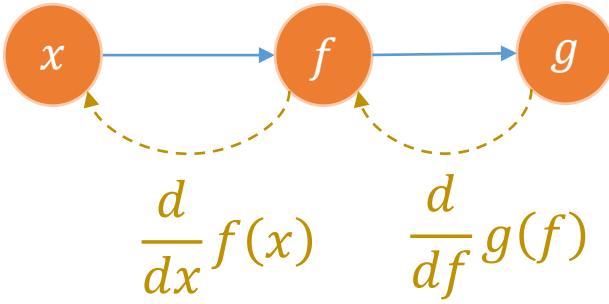
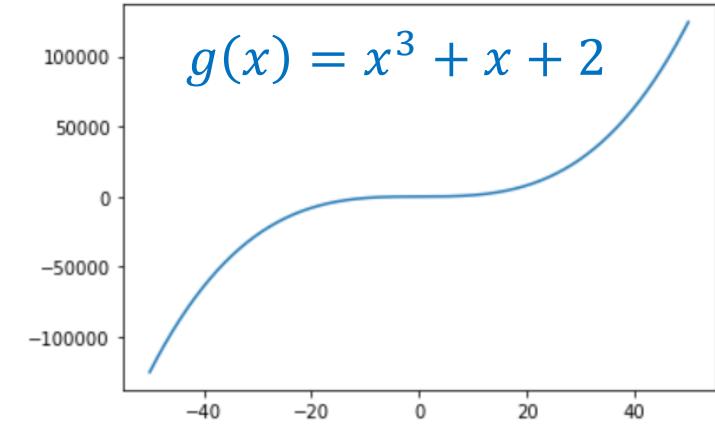
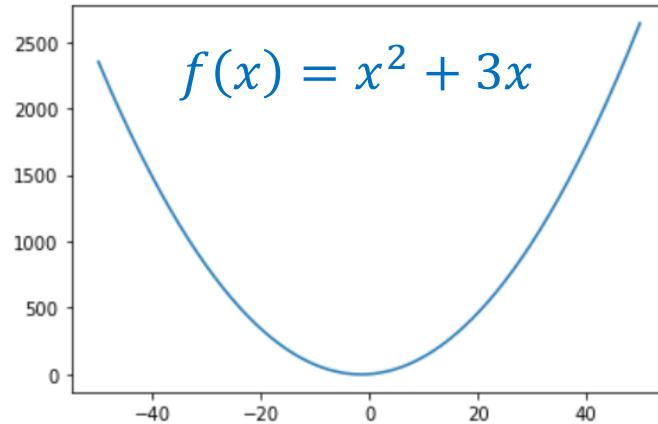
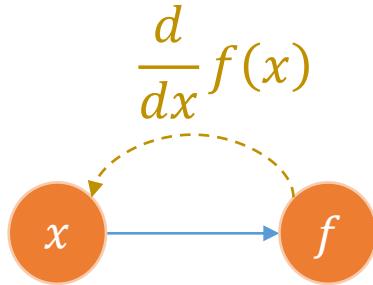
$$x_t = x_{t-1} - \eta f'(x)$$

Observation?



Gradient-based Optimization

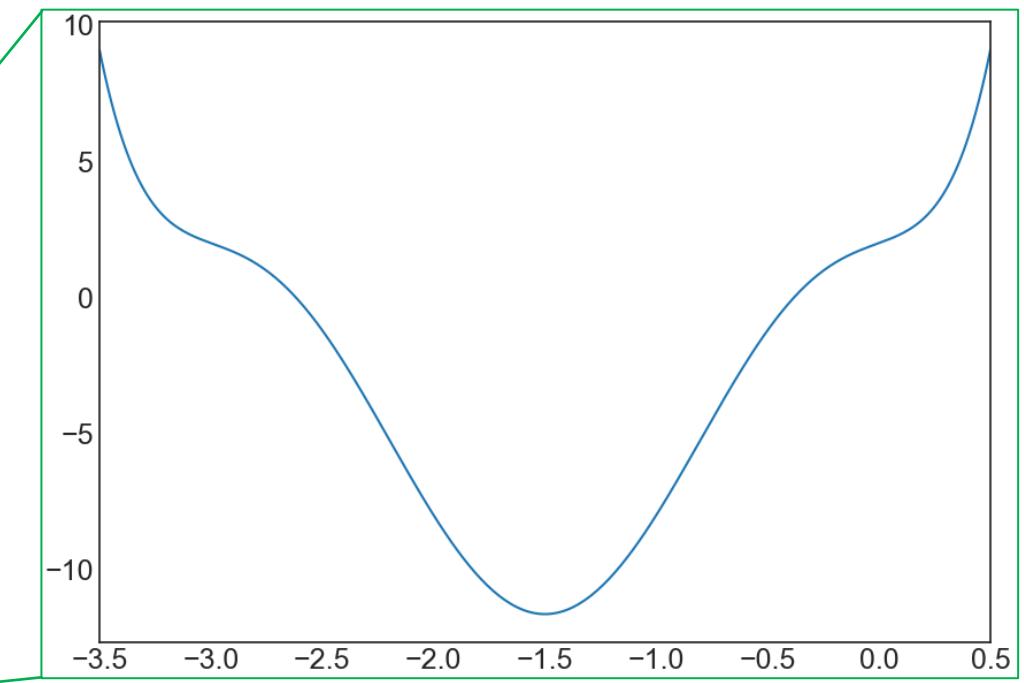
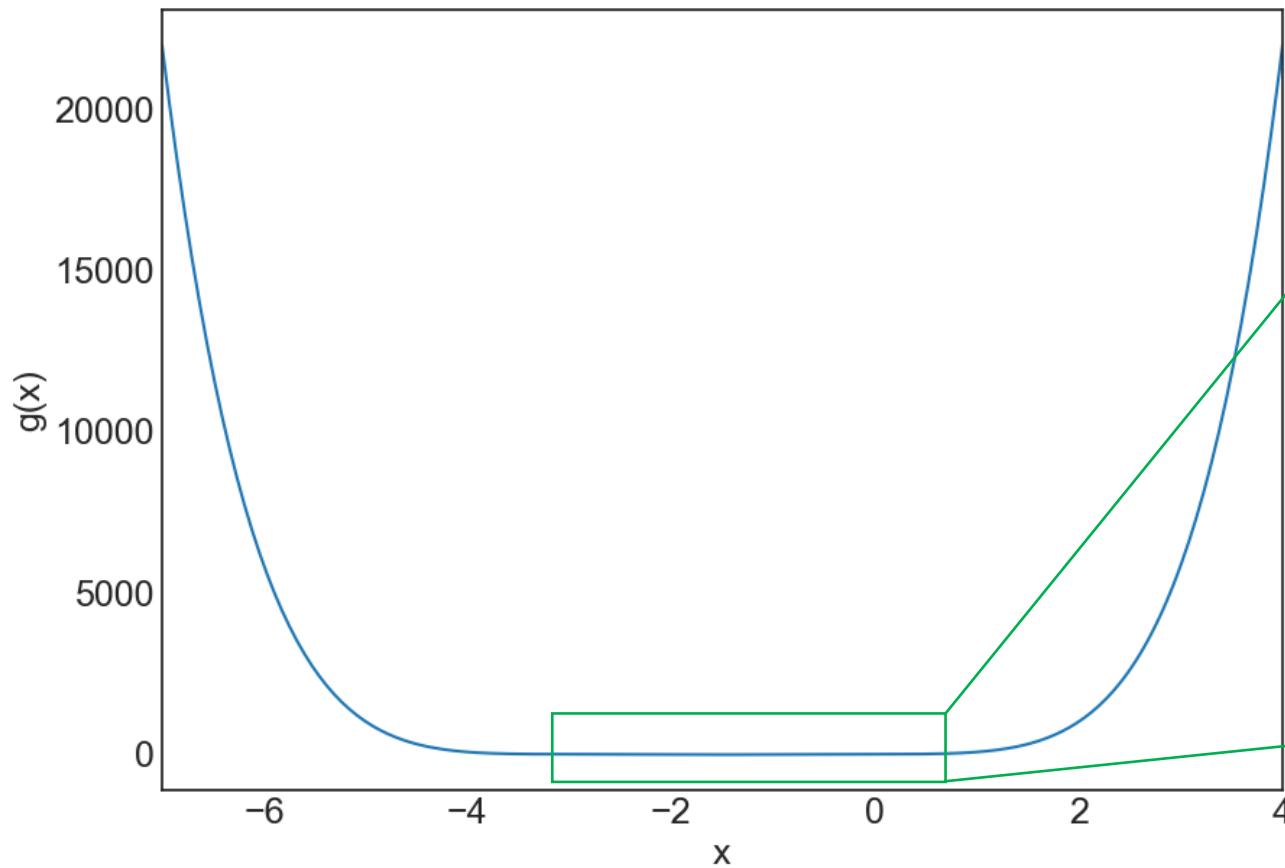
❖ For composite function



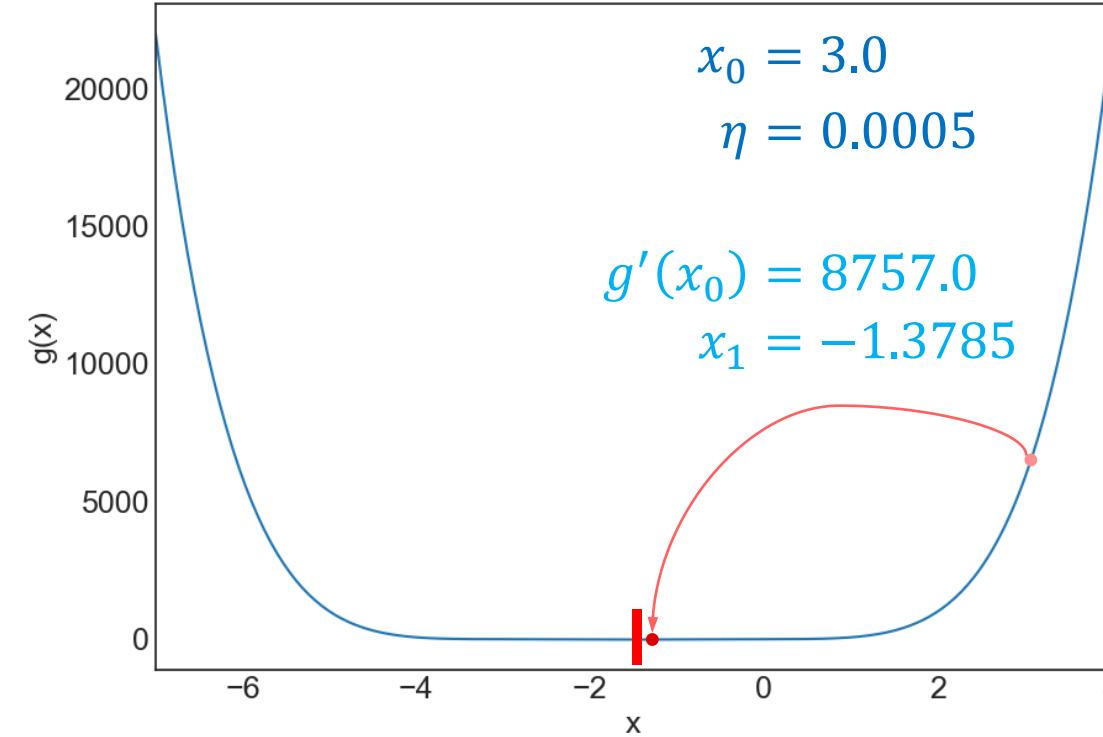
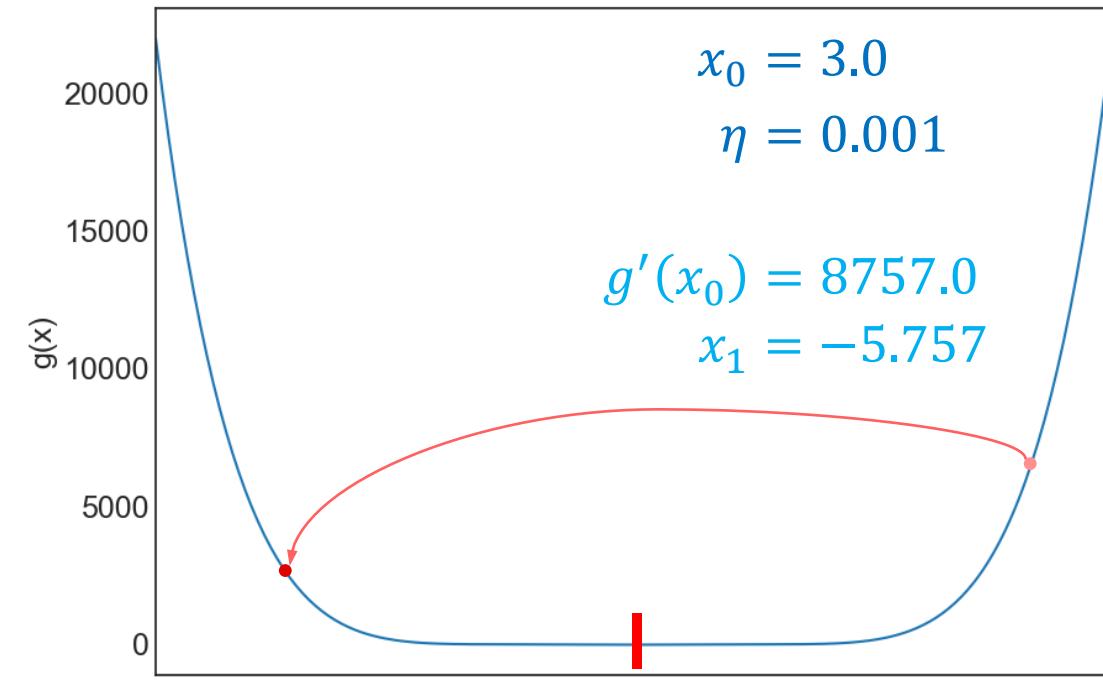
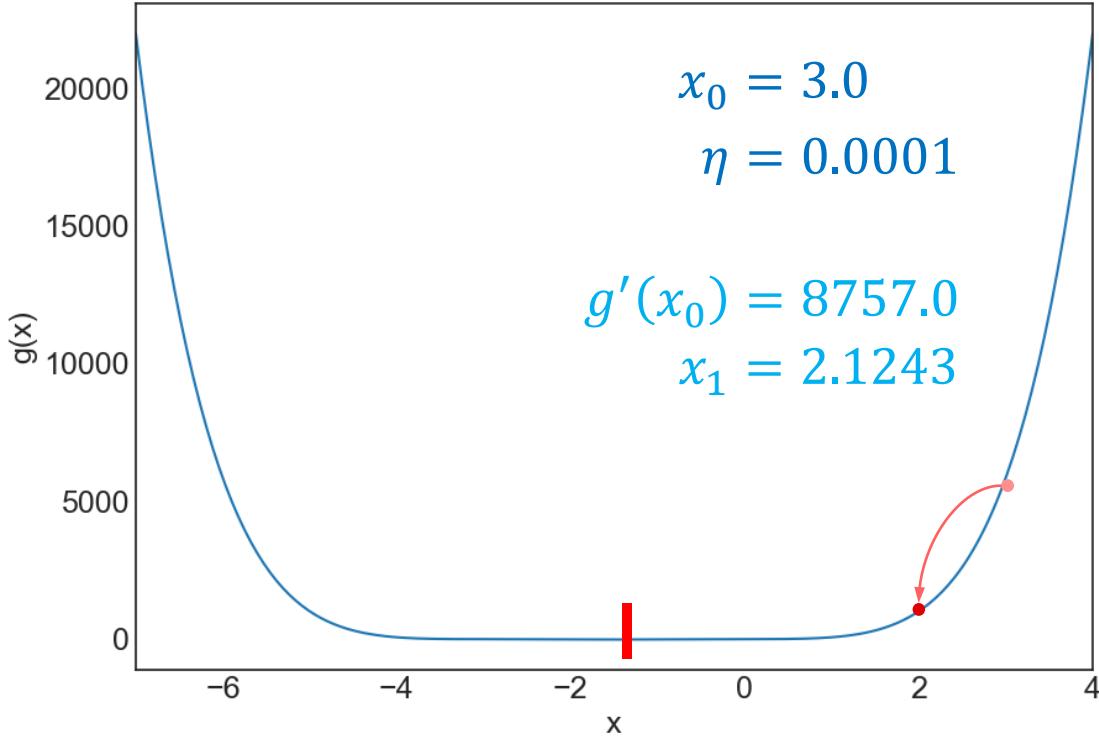
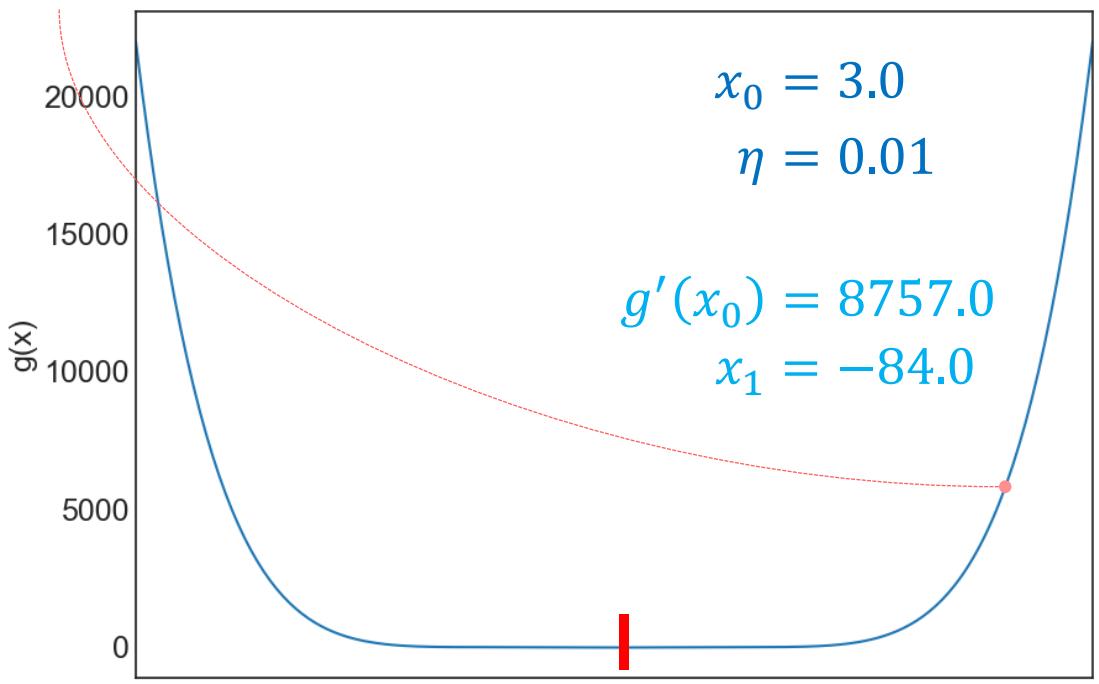
$$\frac{d}{dx} g(f(x)) = \left[\frac{d}{df} g(f) \right] * \left[\frac{d}{dx} f(x) \right]$$

Gradient-based Optimization

❖ For composite function



Select an appropriate
value for learning rate



Optimization

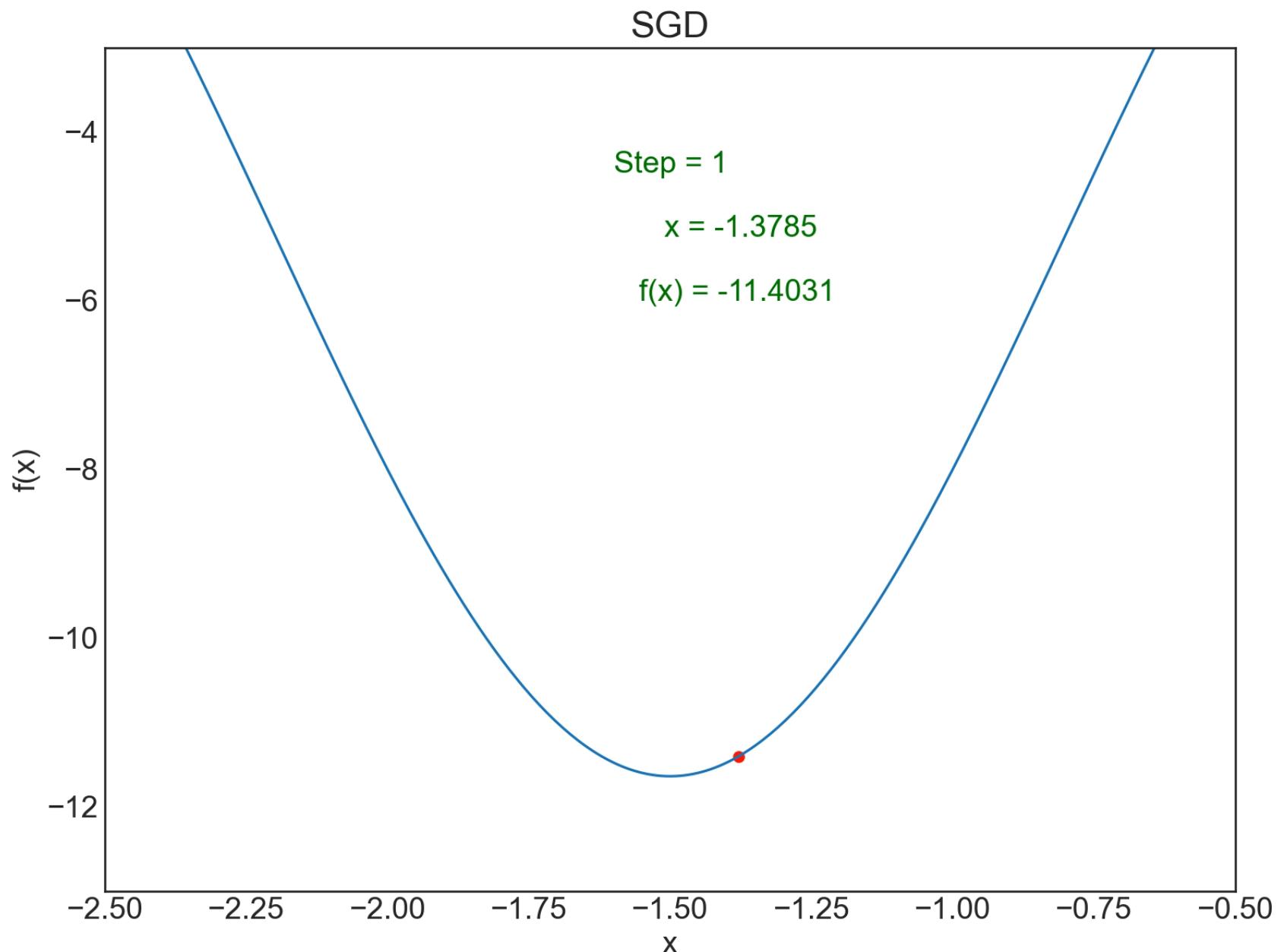
❖ Composite function

$$x_0 = 3.0$$

$$\eta = 0.0005$$

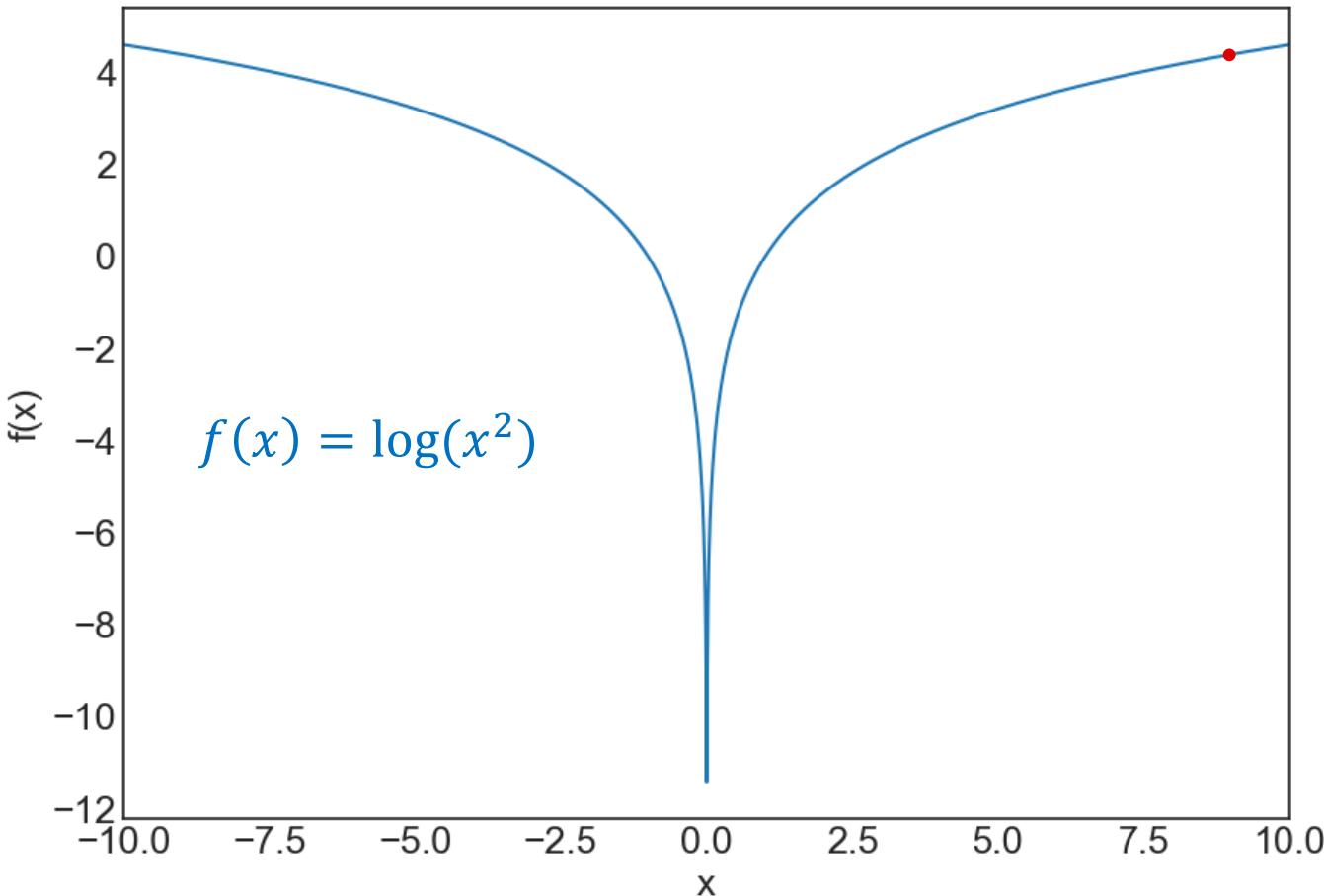
$$x_t = x_{t-1} - \eta f'(x)$$

Observation?



Gradient-based Optimization

❖ Another function



$$x_t = x_{t-1} - \eta f'(x_{t-1})$$

$$x_0 = 9.0 \quad \eta = 10.0$$

$$f'(x_0) = 0.22$$

$$x_1 = x_0 - \eta f'(x_0) = 6.77$$

$$f'(x_1) = 0.295$$

$$x_2 = x_1 - \eta f'(x_1) = 3.82$$

$$f'(x_2) = 0.52$$

$$x_3 = x_2 - \eta f'(x_2) = -1.39$$

$$f'(x_3) = -1.429$$

$$x_4 = x_3 - \eta f'(x_3) = 12.89$$

Optimization

❖ Another function

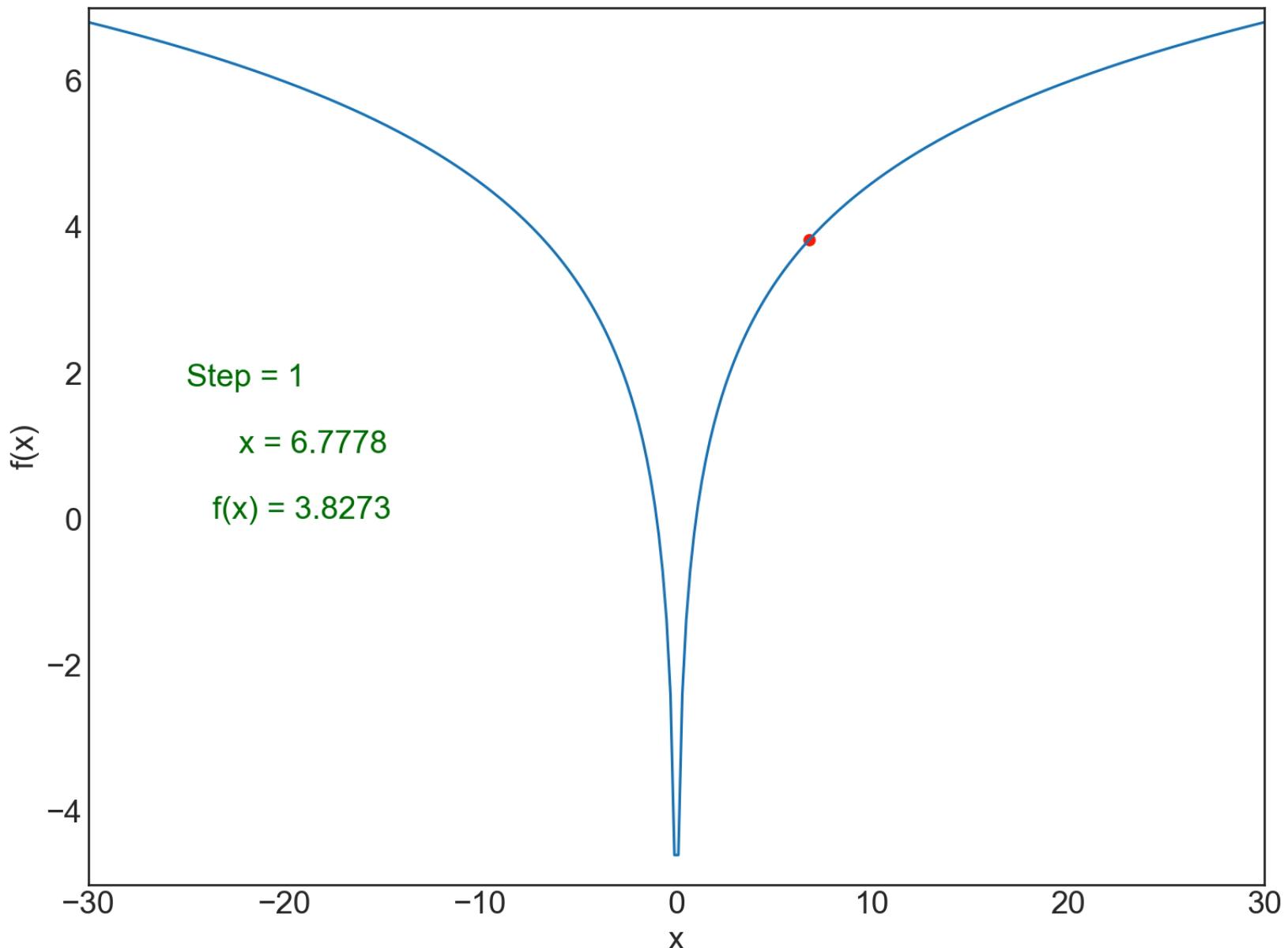
$$x_0 = 9.0$$

$$\eta = 10.0$$

$$x_t = x_{t-1} - \eta f'(x)$$

Observation?

SGD



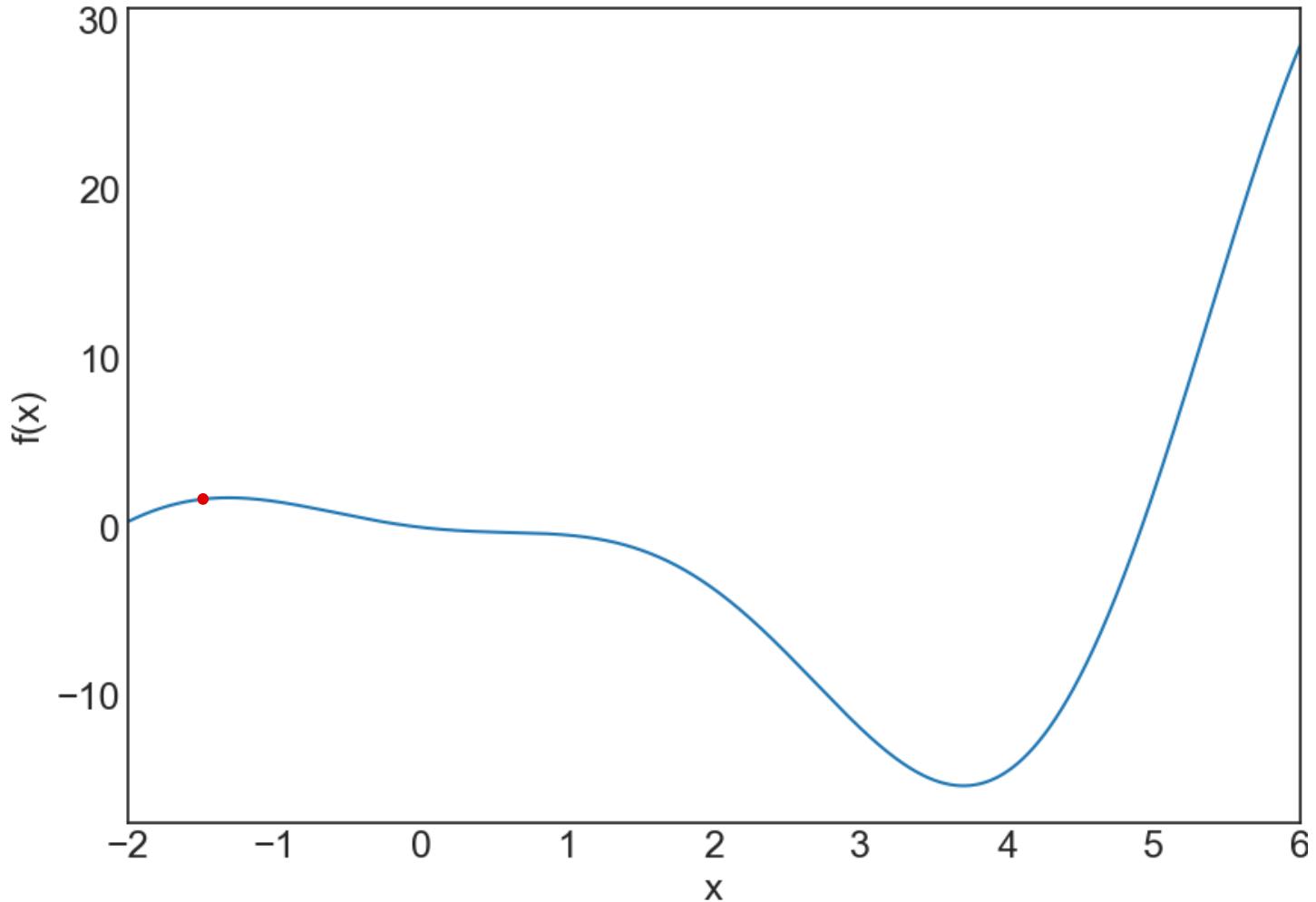
Gradient-based Optimization

❖ Another function

$$x_0 = -1.5$$

$$\eta = 0.2$$

$$x_t = x_{t-1} - \eta f'(x)$$



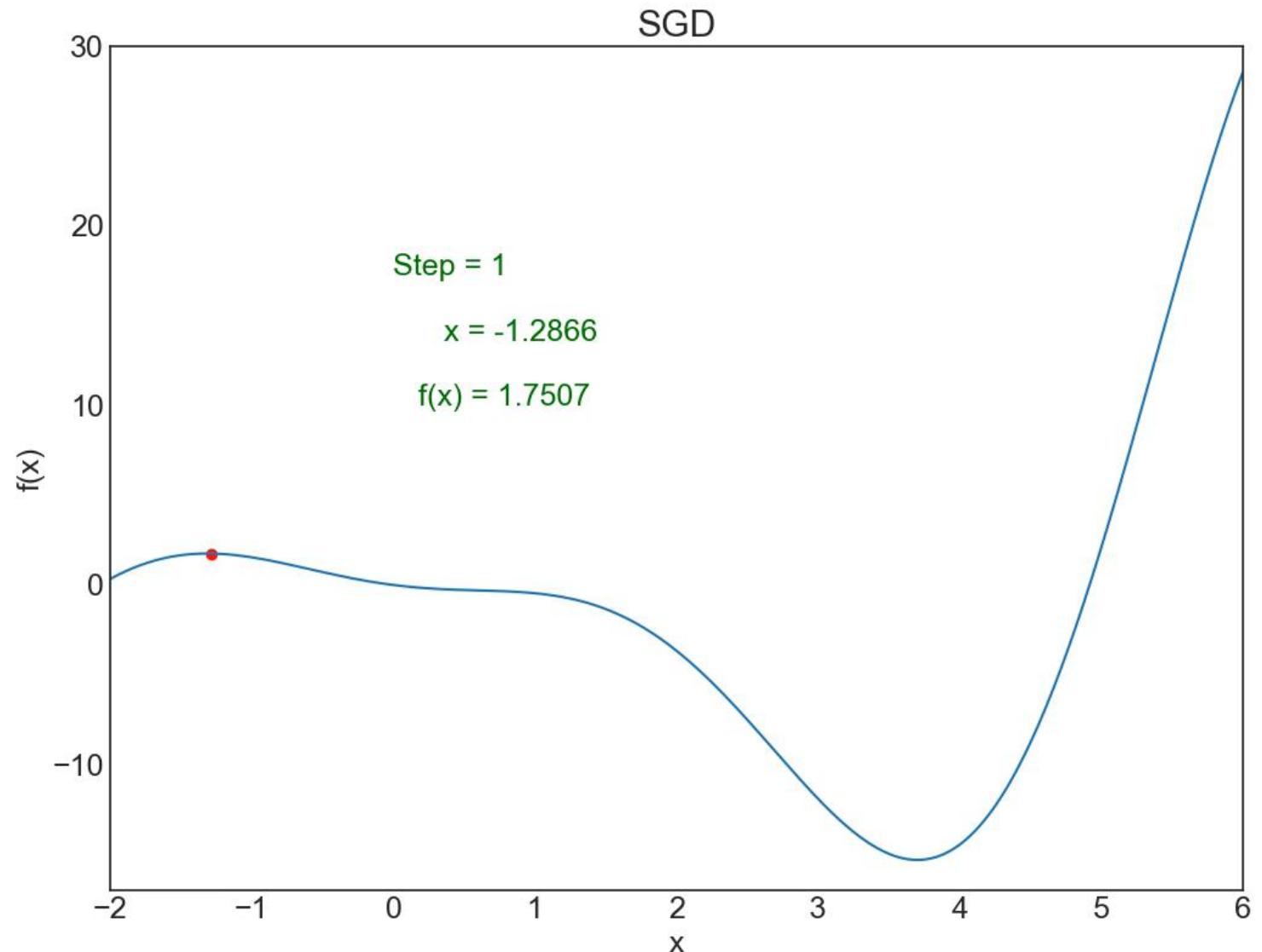
Gradient-based Optimization

❖ Another function

$$x_0 = -1.5$$

$$\eta = 0.2$$

$$x_t = x_{t-1} - \eta f'(x)$$

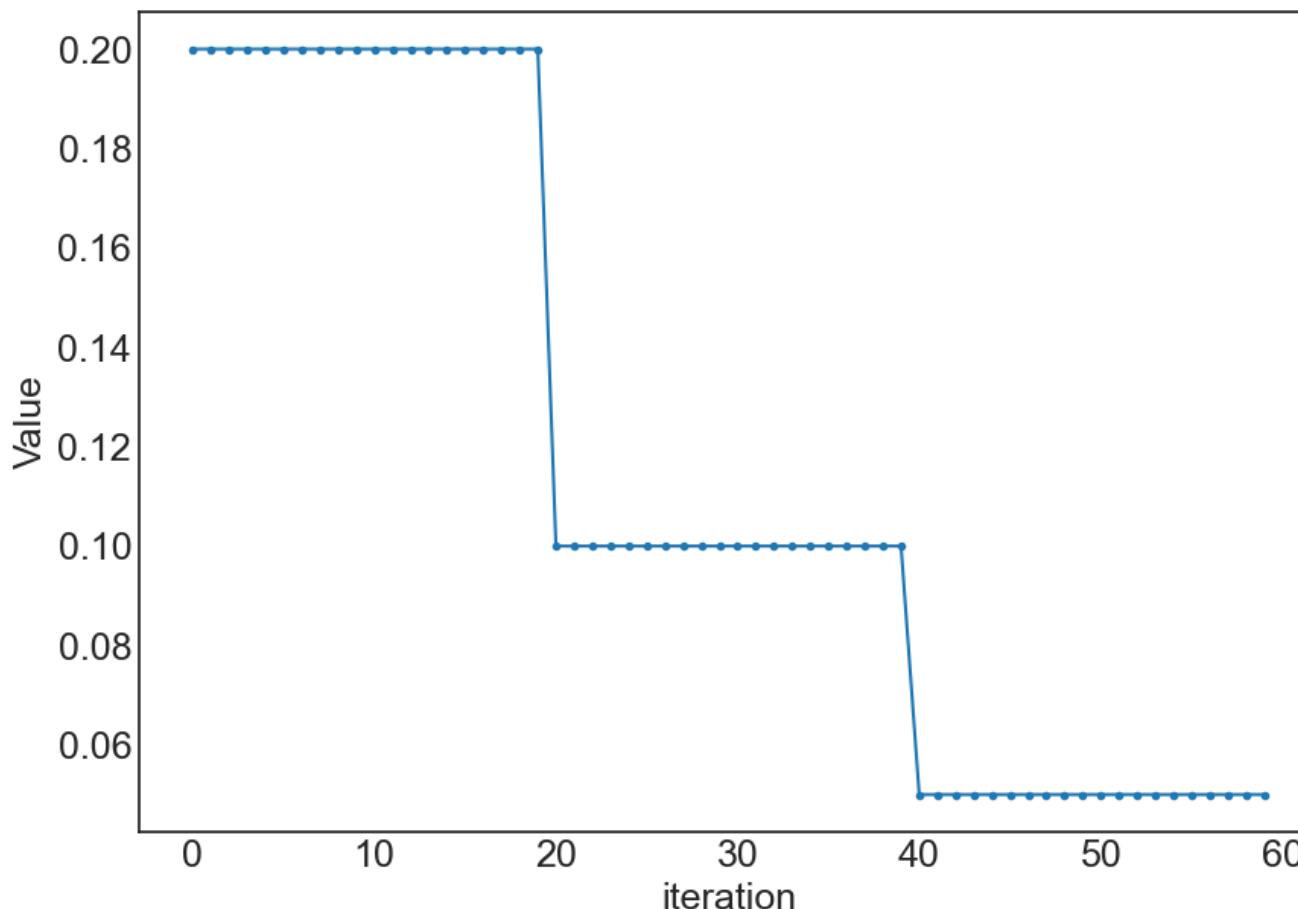


Outline

- Quick Introduction
- Stochastic Gradient Descent
- Adagrad
- RMSProp
- Momentum
- Adam
- Case studies

Adaptive Learning Rate

Learning rate decay



Iteration i

Period p

Reduction rate k

$\eta = 0.2$

$if\ i \% p == 0:$

$\eta = \eta / k$

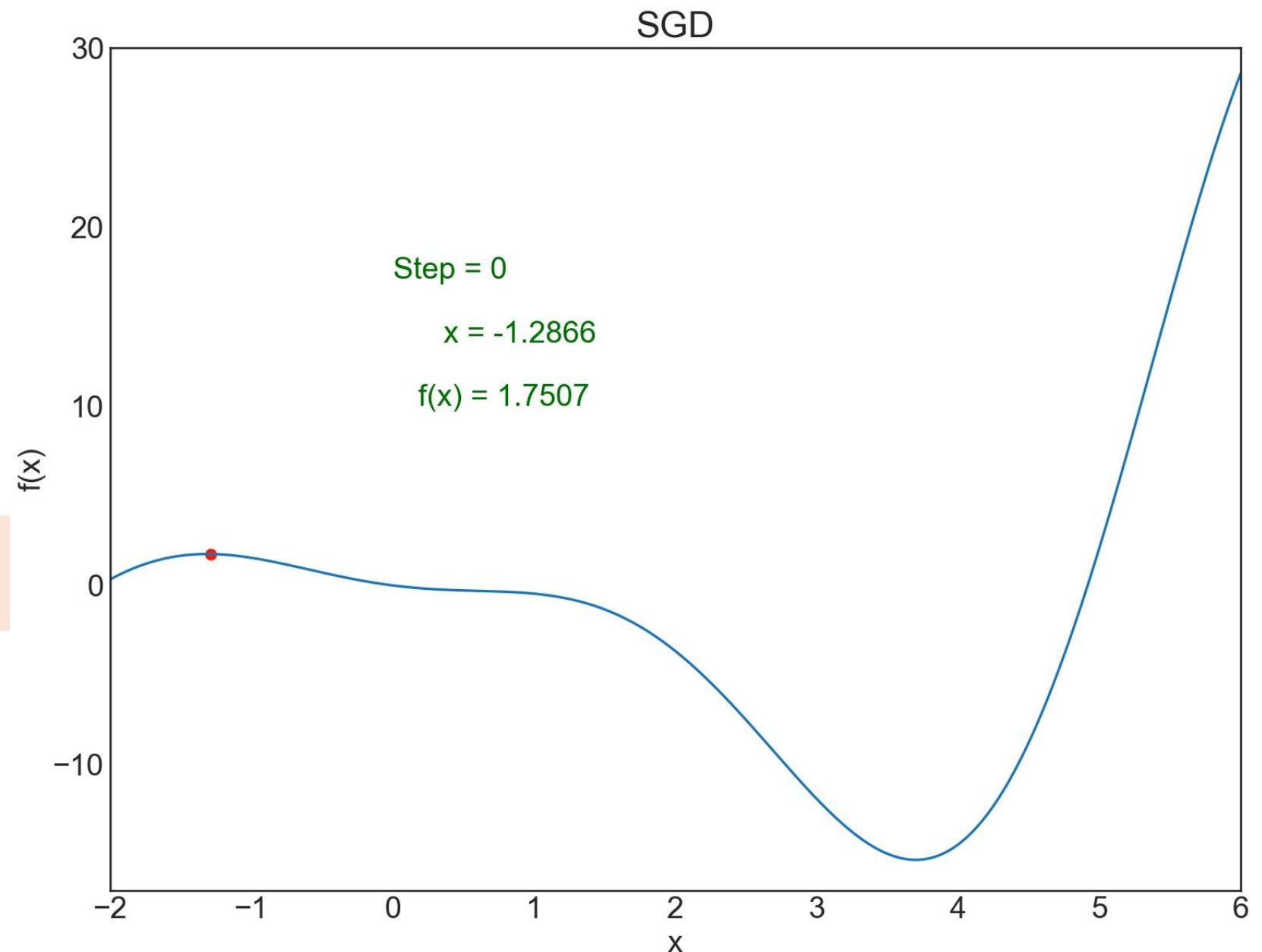
Adaptive Learning Rate

Learning rate decay

$$x_0 = -1.5$$

$$\eta = 0.2$$

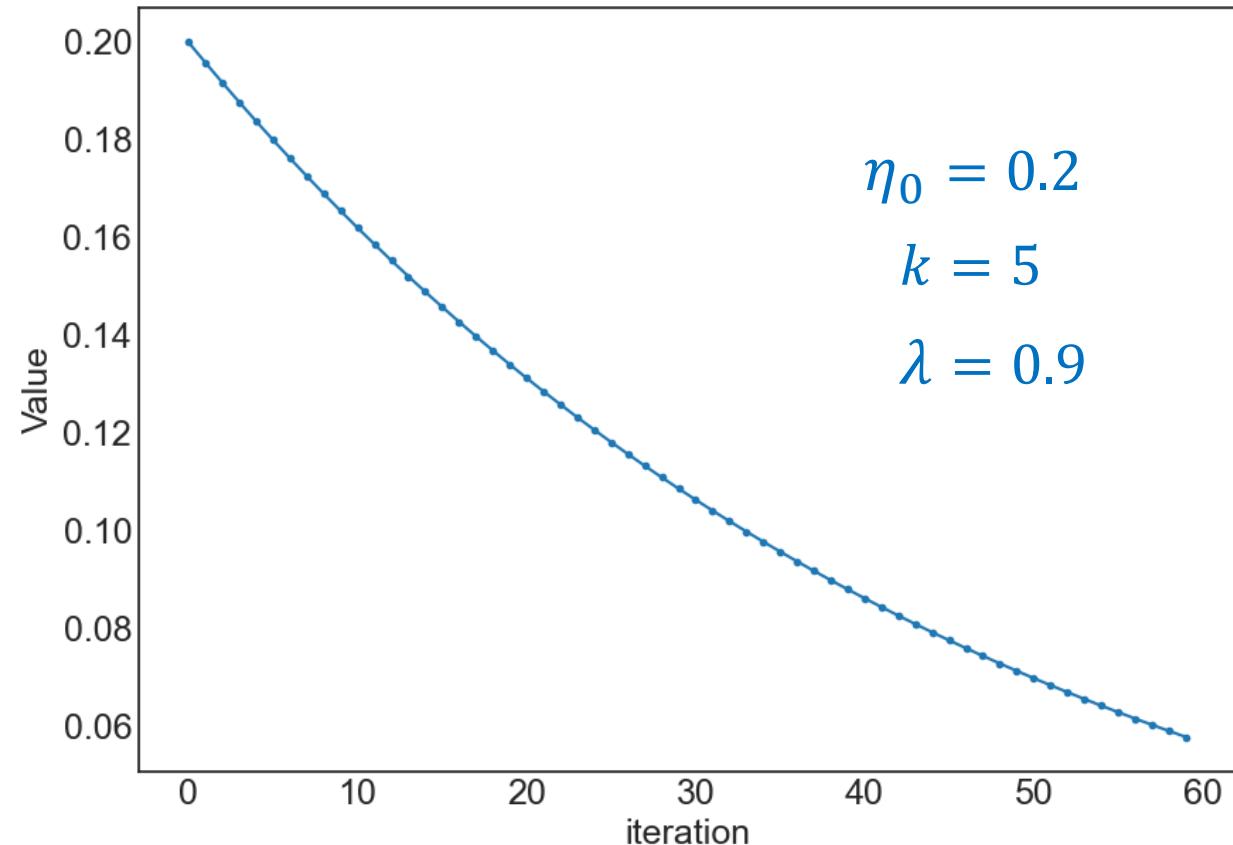
$$x_t = x_{t-1} - \eta f'(x)$$



Adaptive Learning Rate

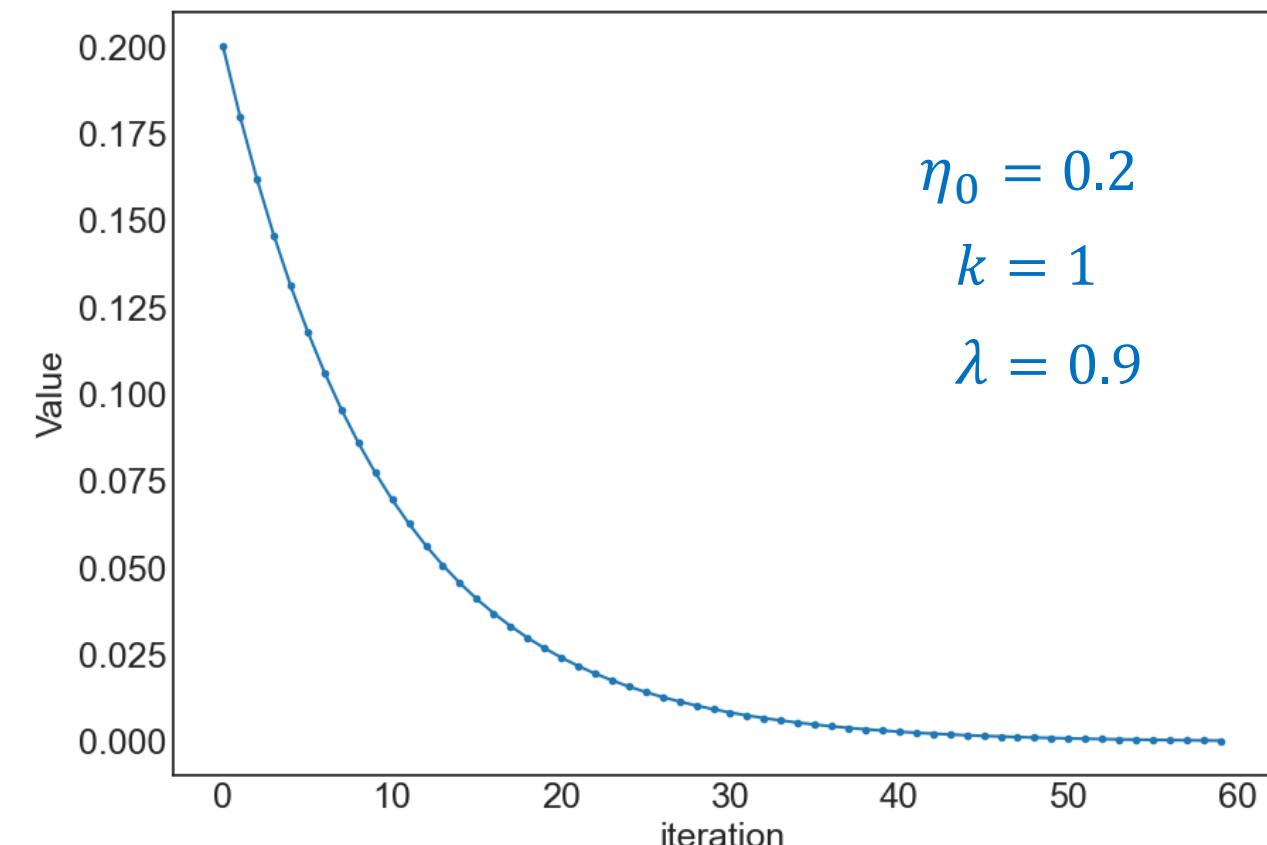
Learning rate decay

$$\eta = \eta_0 \times \lambda^{\frac{s}{k}}$$



`torch.optim.lr_scheduler`

`initial_learning_rate * decay_rate ^ (step / decay_steps)`



Adaptive Learning Rate

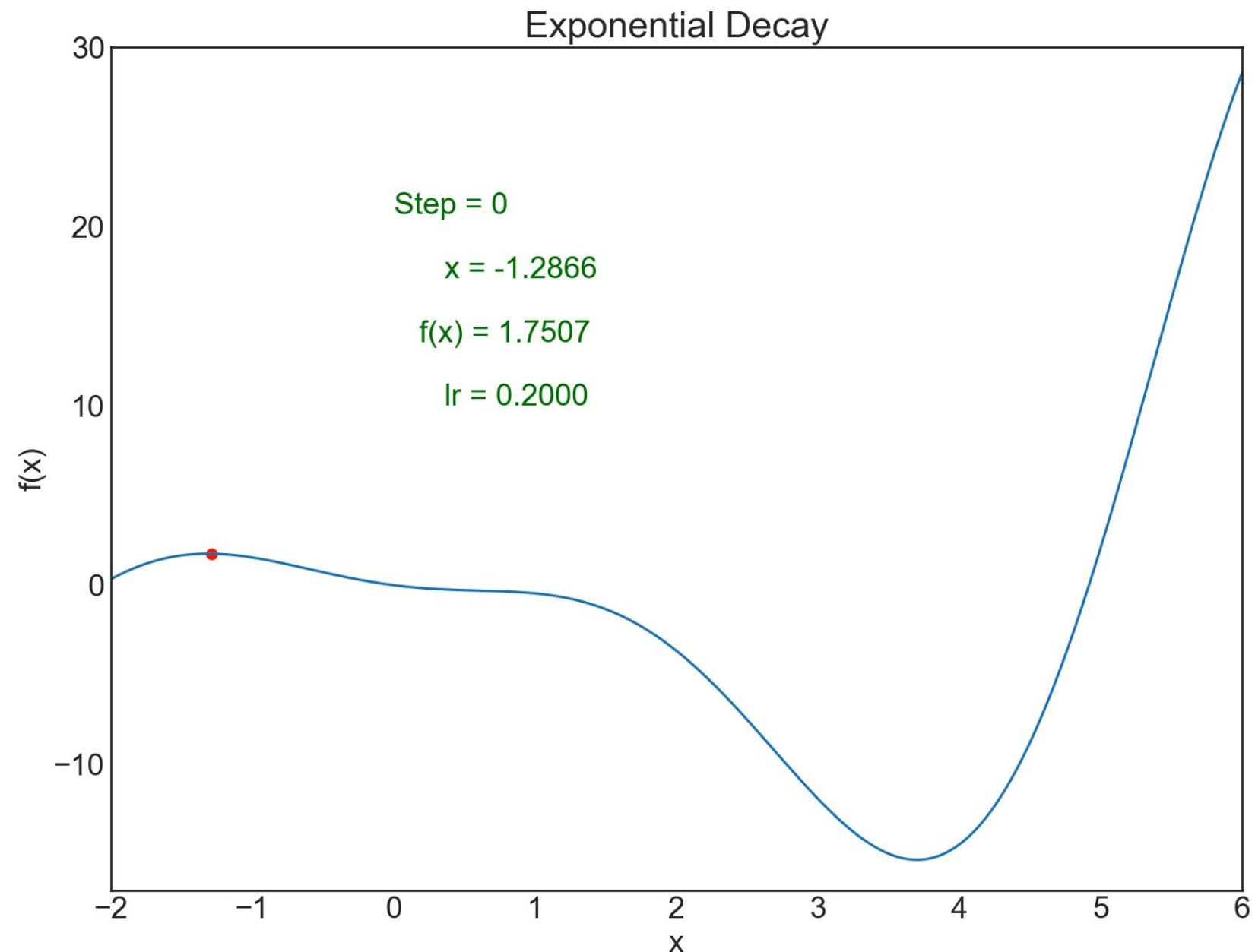
Learning rate decay

$$x_0 = -1.5$$

$$\eta = 0.2$$

$$x_t = x_{t-1} - \eta f'(x)$$

$$\eta = \eta_0 \times \lambda^{\frac{s}{k}}$$



Adaptive Learning Rate

Using derivative values

Adagrad (one variable functions)

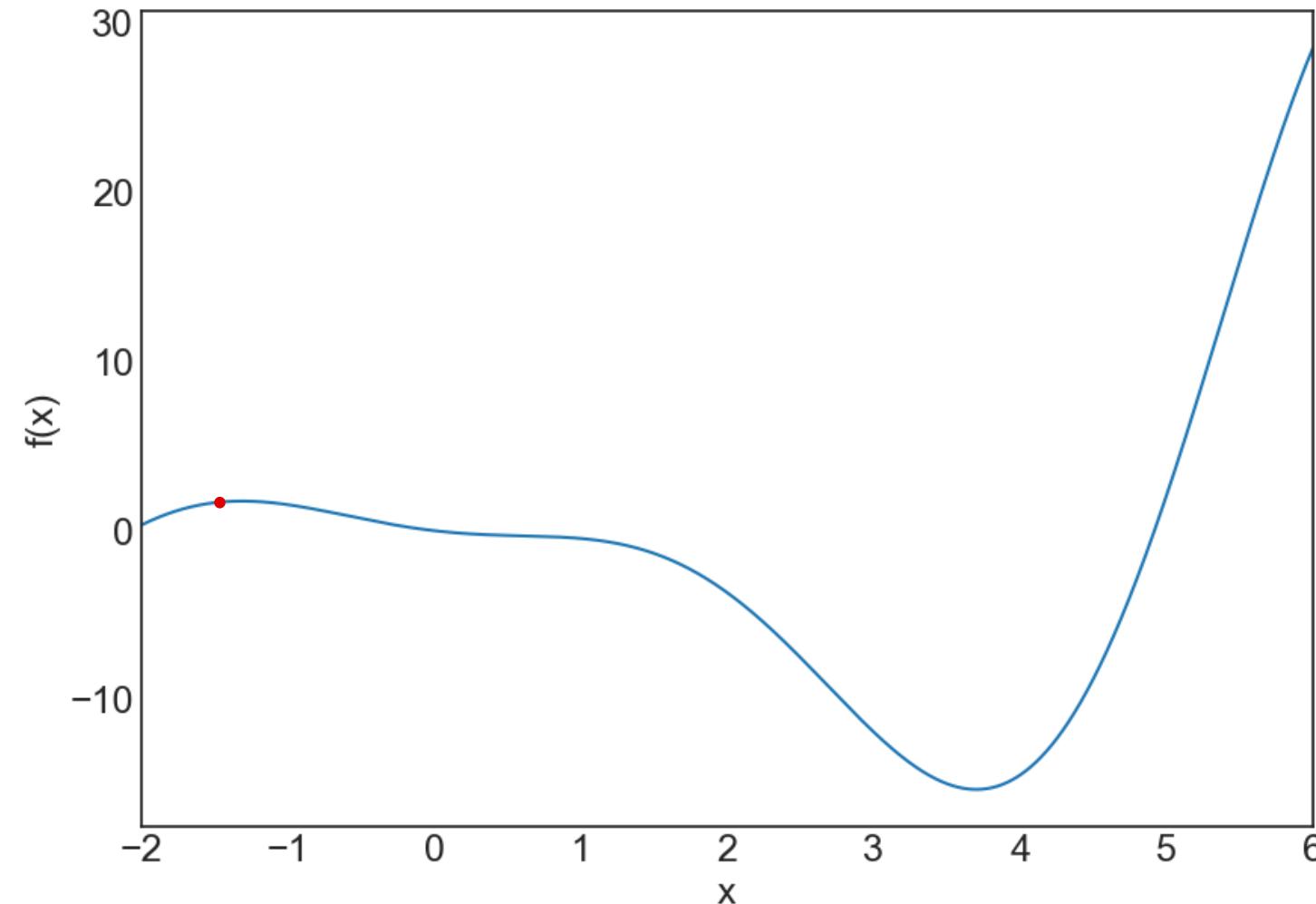
$$s_0 = 0.1 \text{ (or } s_0 = 0.0\text{)}$$

$$\epsilon = 10^{-7}$$

$$g_t = f'(x_{t-1})$$

$$s_t = s_{t-1} + g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$



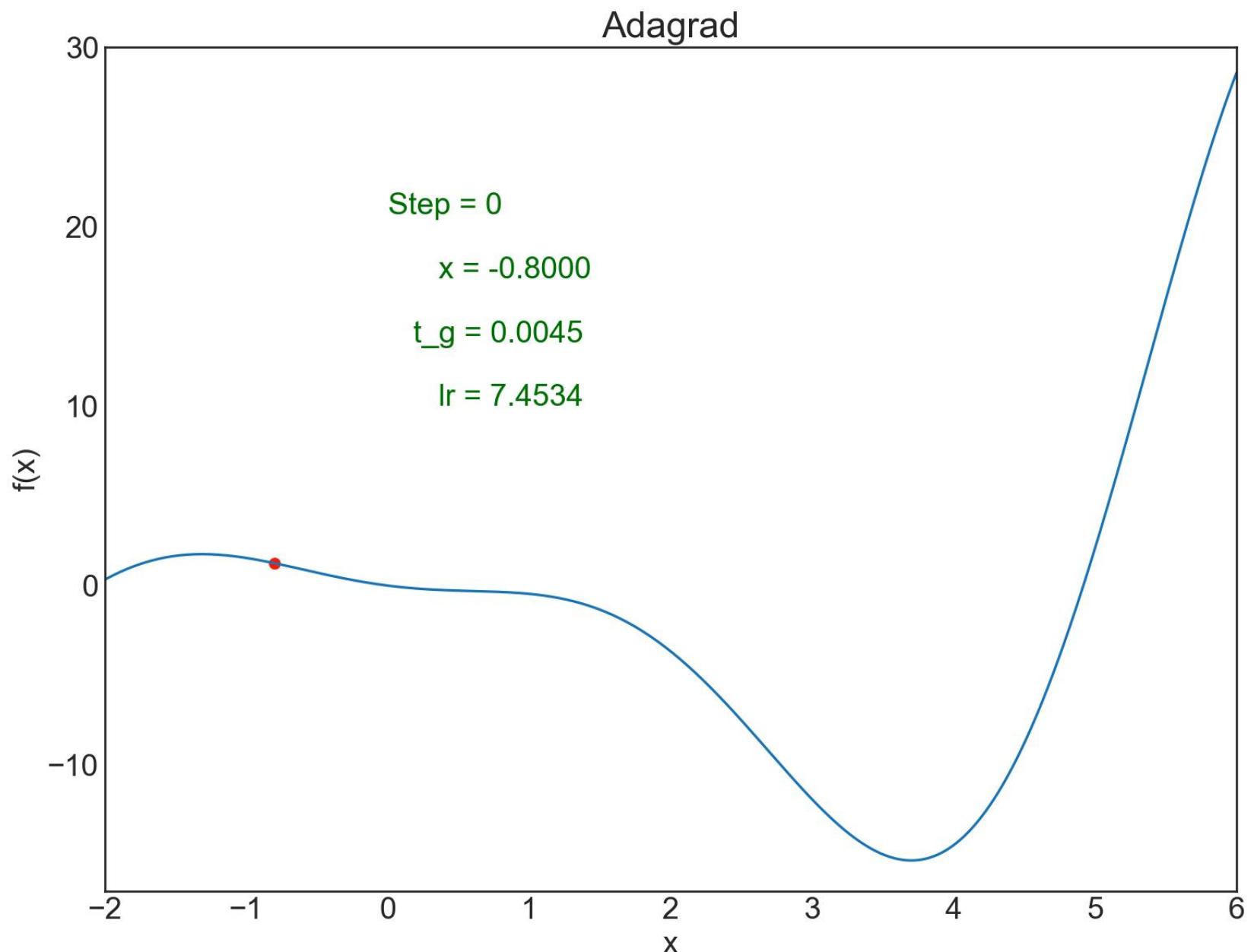
Adaptive Learning Rate

Adagrad
(one variable functions)

$$\eta = 0.5$$

$$s_0 = 0.0$$

$$\epsilon = 10^{-7}$$



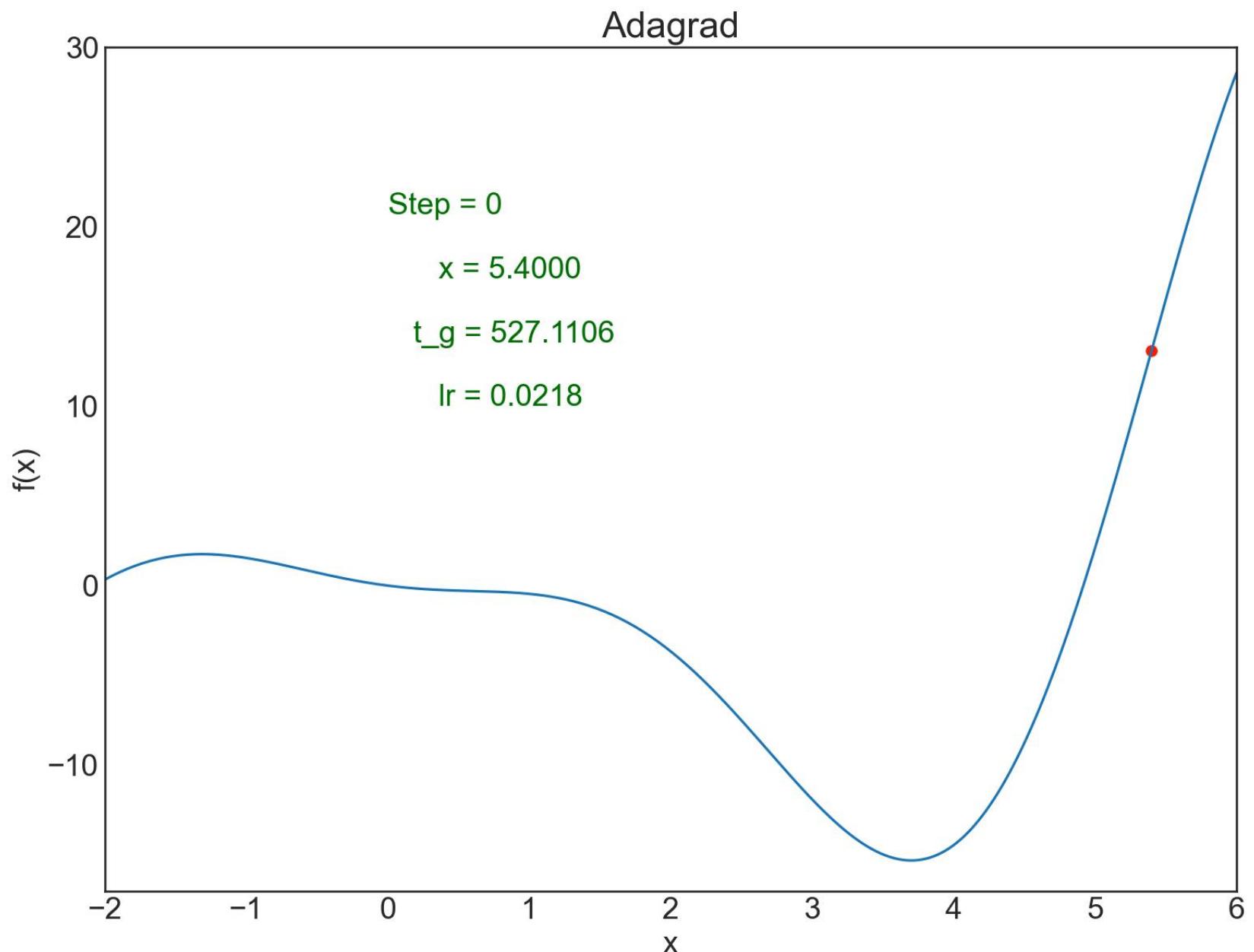
Adaptive Learning Rate

Adagrad
(one variable functions)

$$\eta = 0.5$$

$$s_0 = 0.0$$

$$\epsilon = 10^{-7}$$



Adaptive Learning Rate

Using derivative values

Adagrad (one variable functions)

$$s_0 = 0.1$$

$$\epsilon = 10^{-7}$$

$$g_t = f'(x_{t-1})$$

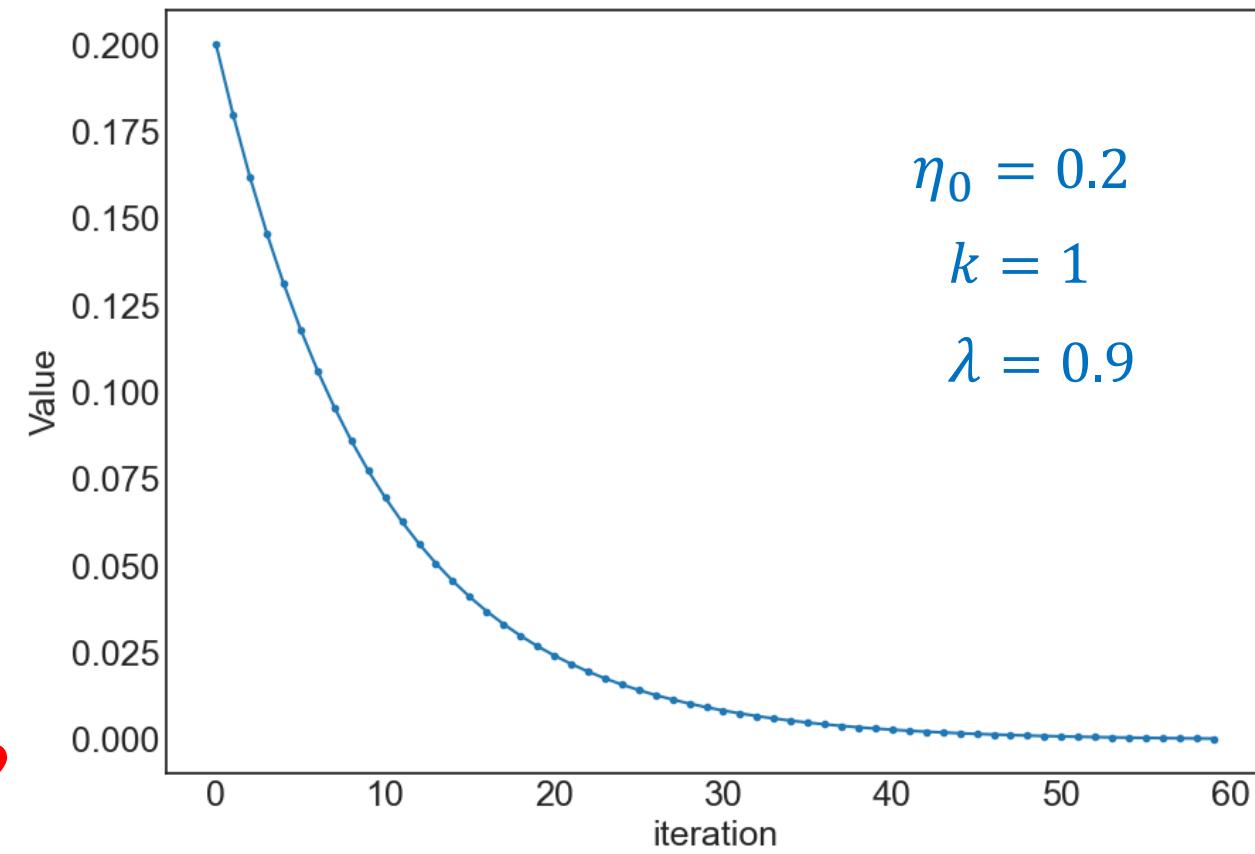
$$s_t = s_{t-1} + g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

Differences?

Learning rate decay

$$\eta = \eta_0 \times \lambda^{\frac{s}{k}}$$



Adaptive Learning Rate

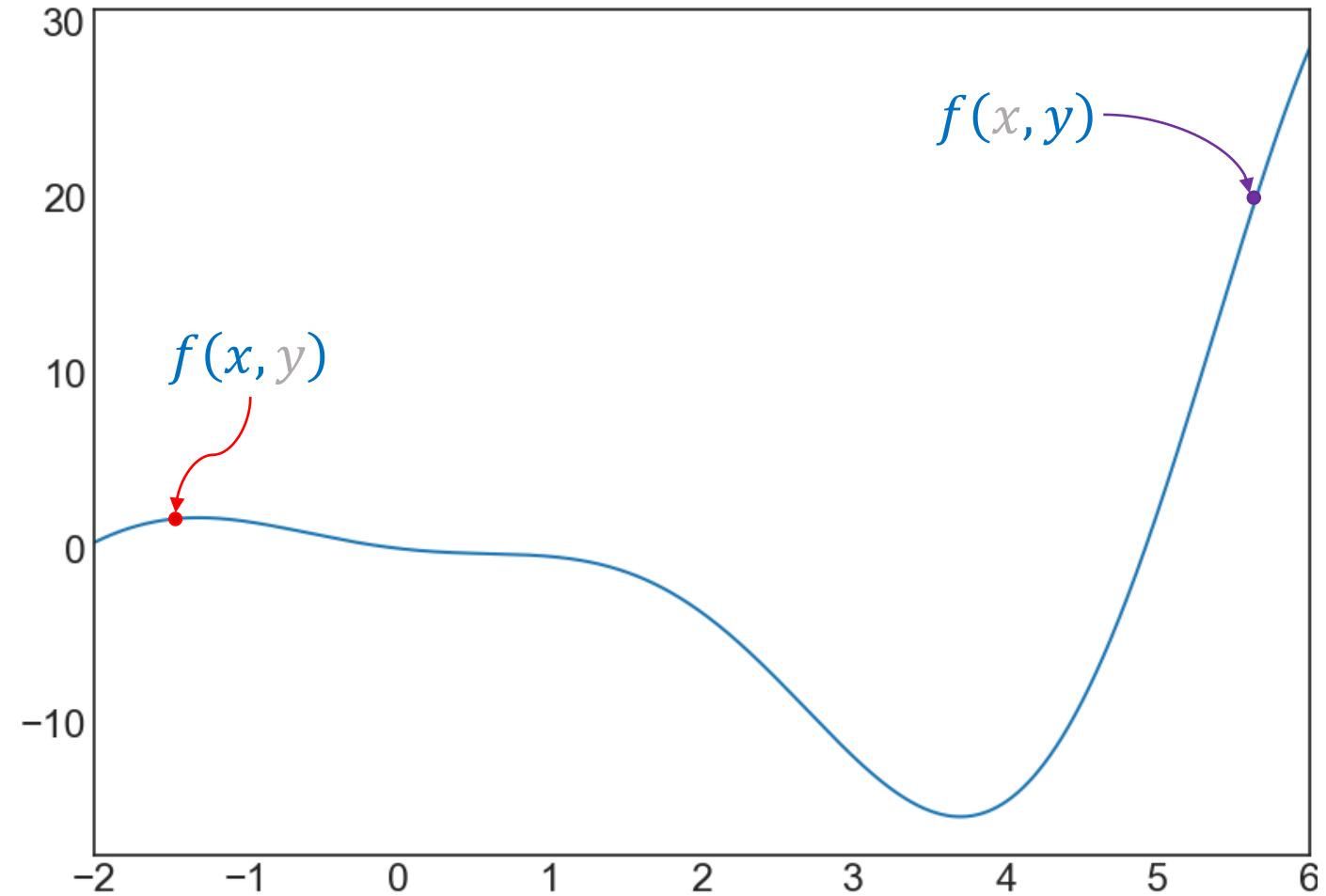
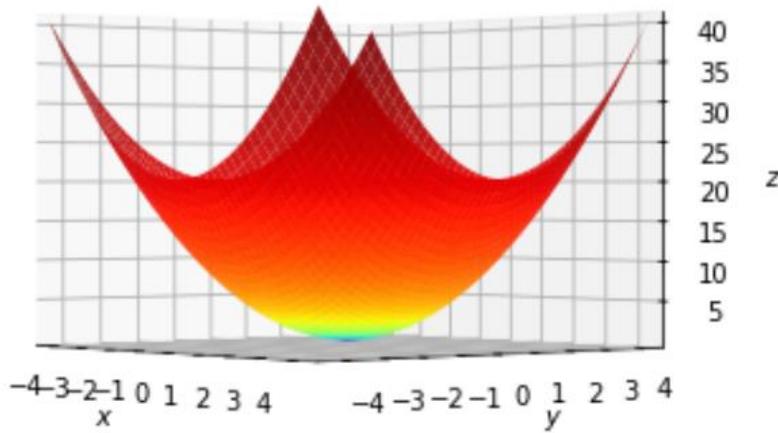
❖ Optimization: 2D function

imagine

$$f(x, y) = x^2 + y^2$$

$$-100 \leq x, y \leq 100$$

$$x, y \in \mathbb{N}$$



Adaptive Learning Rate

Adagrad (2D function)

$$g_{t,x} = \frac{\partial f(x, y)}{\partial x}$$

$$s_{t,x} = s_{t-1,x} + g_{t,x}^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_{t,x} + \epsilon}} g_{t,x}$$

$$g_{t,y} = \frac{\partial f(x, y)}{\partial y}$$

$$s_{t,y} = s_{t-1,y} + g_{t,y}^2$$

$$y_t = y_{t-1} - \frac{\eta}{\sqrt{s_{t,y} + \epsilon}} g_{t,y}$$

Adagrad

$$g_t = \nabla_{\theta} L$$

$$s_t = s_{t-1} + g_t^2$$

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

Adaptive Learning Rate

Adagrad: Limitation

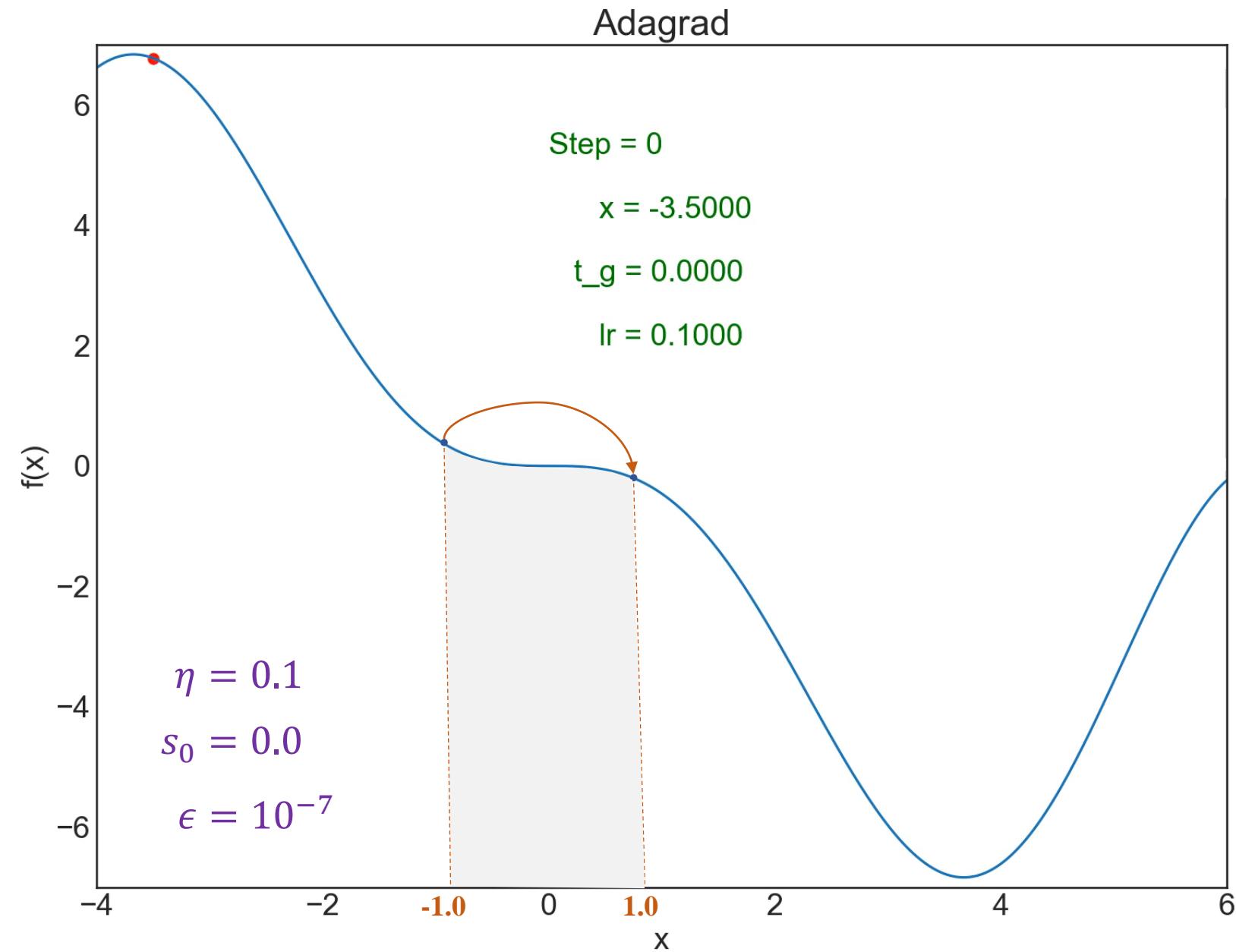
$$g_t = \nabla_{\theta} L$$

$$s_t = s_{t-1} + g_t^2$$

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

current summation

expected summation



How to Use historical Data

Moving average

$k = 2$

3	8	6	5	1	7	9	0	8	4
---	---	---	---	---	---	---	---	---	---

5.5	7.0	5.5	3.0	4.0	8.0	4.5	4.0	6.0
-----	-----	-----	-----	-----	-----	-----	-----	-----

$$SMA_t = \frac{s_{t-1} + s_{t-2} + \dots + s_{t-k}}{k}$$

3	8	6	5	1	7	9	0	8	4
---	---	---	---	---	---	---	---	---	---

3.0	5.5	5.8	5.4	3.2	5.1	7.0	3.5	5.8	4.9
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

$$EMA_t = \rho EMA_{t-1} + (1 - \rho)s_t$$

How to Use historical Data

Exponentially weighted averages

3	8	6	5	1	7	9	0	8	4
---	---	---	---	---	---	---	---	---	---

3.0	5.5	5.8	5.4	3.2	5.1	7.0	3.5	5.8	4.9
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

$$V_t = \rho V_{t-1} + (1 - \rho) s_t$$

$$V_1 = \rho V_0 + (1 - \rho) s_1$$

$$V_3 = \rho[\rho V_1 + (1 - \rho) s_2] + (1 - \rho) s_3$$

$$V_2 = \rho V_1 + (1 - \rho) s_2$$

$$V_3 = \rho[\rho[\rho V_0 + (1 - \rho) s_1] + (1 - \rho) s_2] + (1 - \rho) s_3$$

$$V_3 = \rho V_2 + (1 - \rho) s_3$$

Given $V_0 = 0$, we have

$$V_3 = \rho[\rho(1 - \rho)s_1 + (1 - \rho)s_2] + (1 - \rho)s_3$$

$$V_3 = \rho^2 (1 - \rho)s_1 + \rho(1 - \rho)s_2 + (1 - \rho)s_3$$

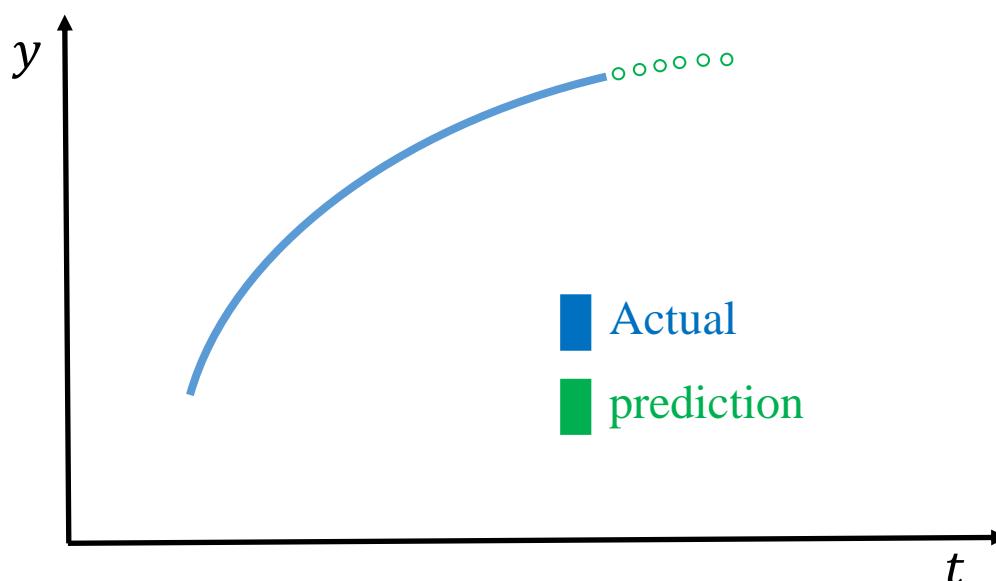
How to Use historical Data

Exponentially weighted averages

$$V_t = \rho V_{t-1} + (1 - \rho)s_t$$

data	3	8	6	5	1	7	9	0	8	4
------	---	---	---	---	---	---	---	---	---	---

EWA	3.0	5.5	5.8	5.4	3.2	5.1	7.0	3.5	5.8	4.9
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----



Example

$$V_3 = \rho^2 (1 - \rho)s_1 + \rho(1 - \rho)s_2 + (1 - \rho)s_3$$

With $\rho = 0.5$

$$V_3 = 0.125s_1 + 0.25s_2 + 0.5s_3$$

With $\rho = 0.9$

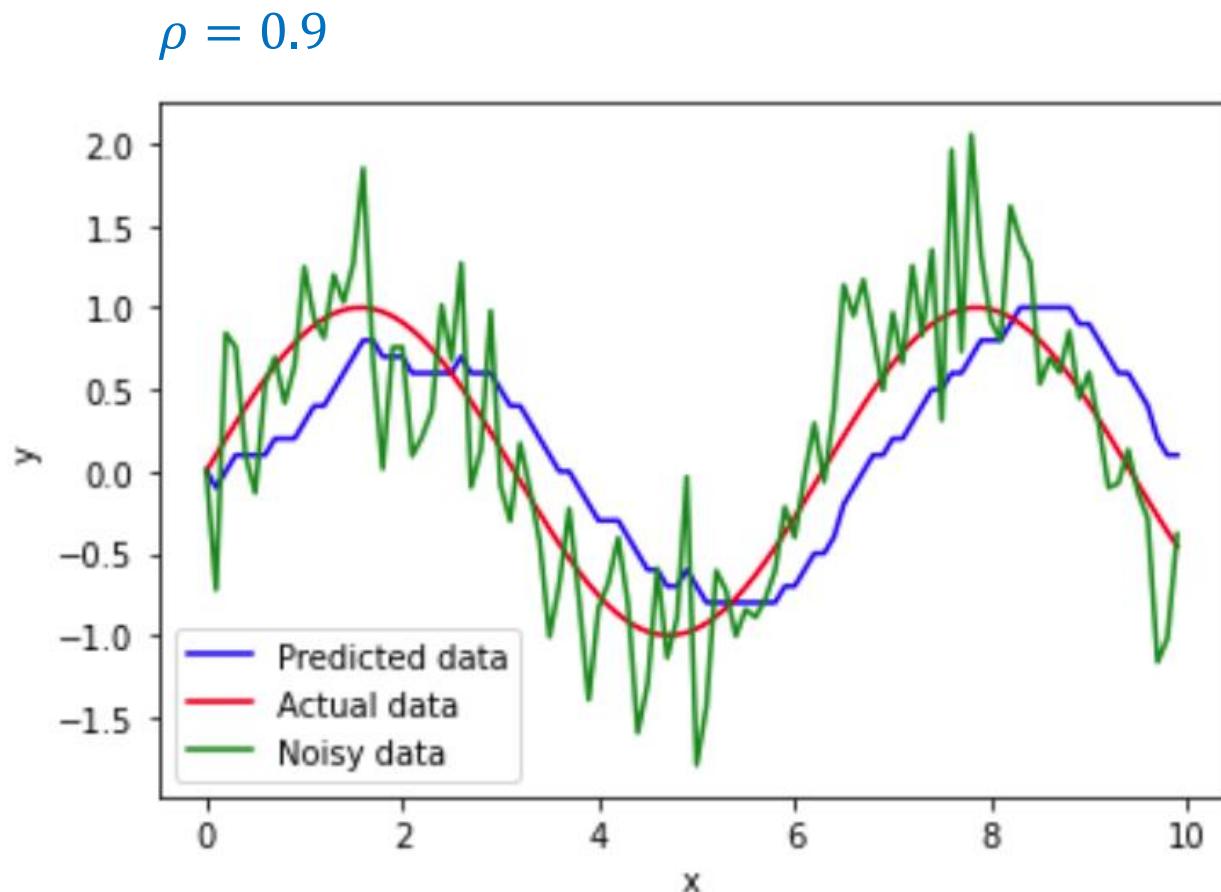
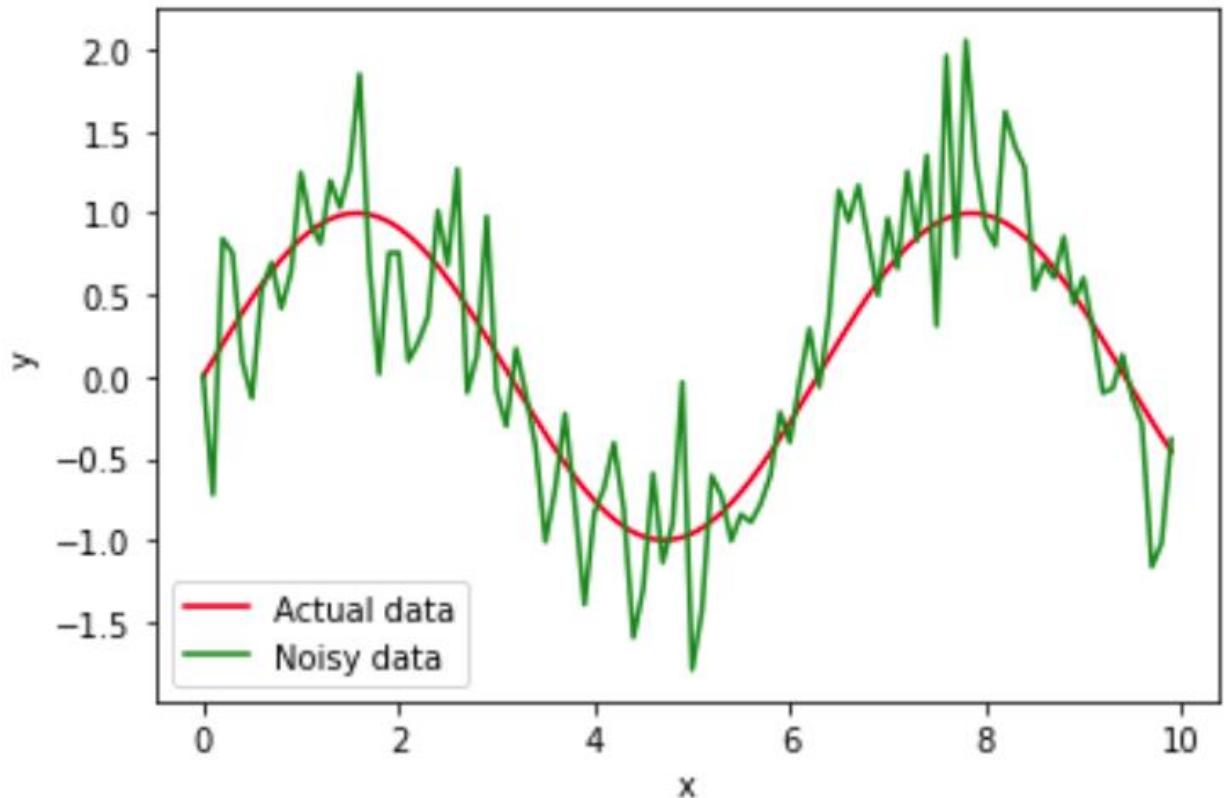
$$V_3 = 0.081s_1 + 0.09s_2 + 0.1s_3$$

With $\rho = 0.98$

$$V_3 = 0.0392s_1 + 0.0196s_2 + 0.02s_3$$

Optimization Algorithms

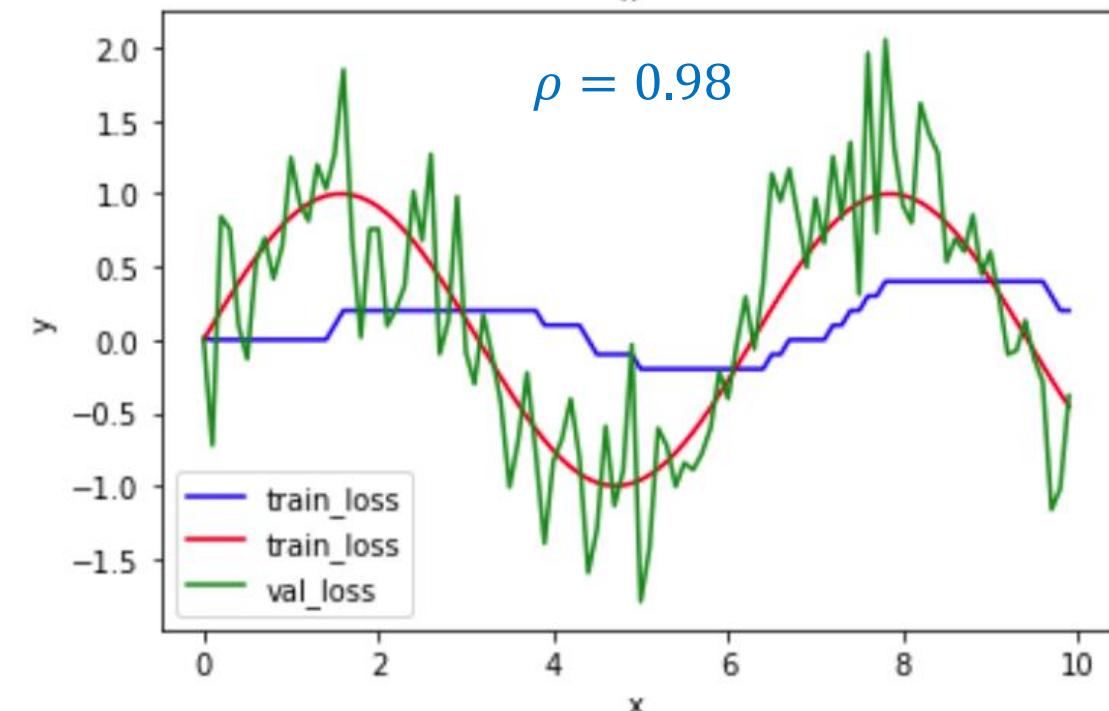
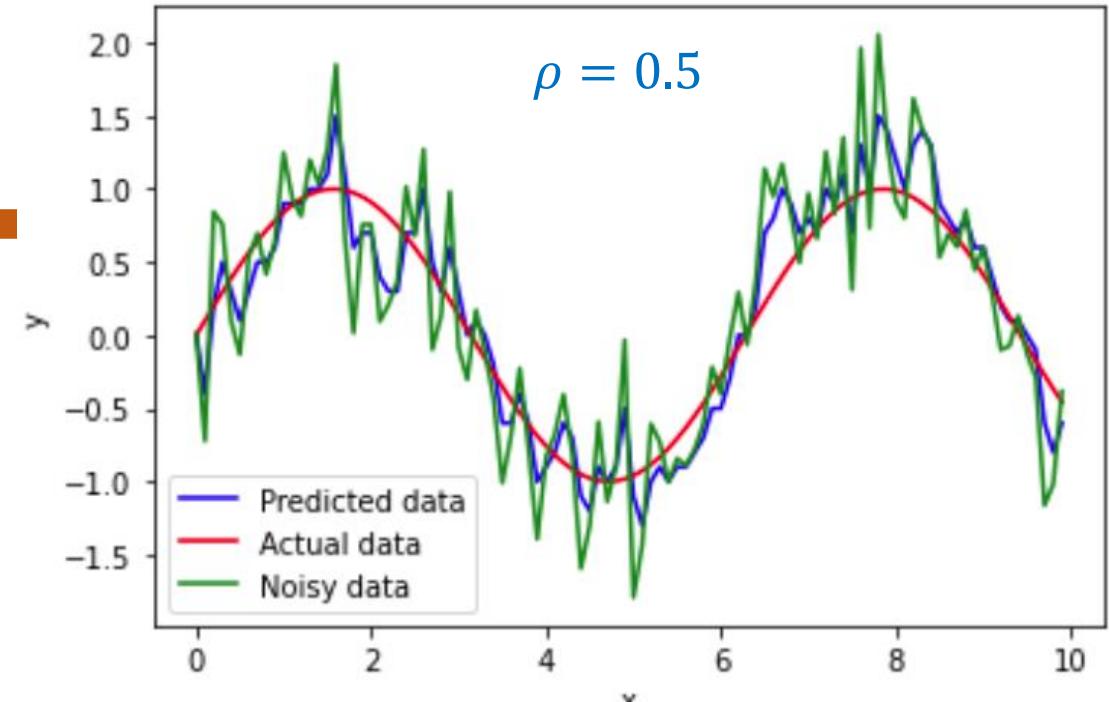
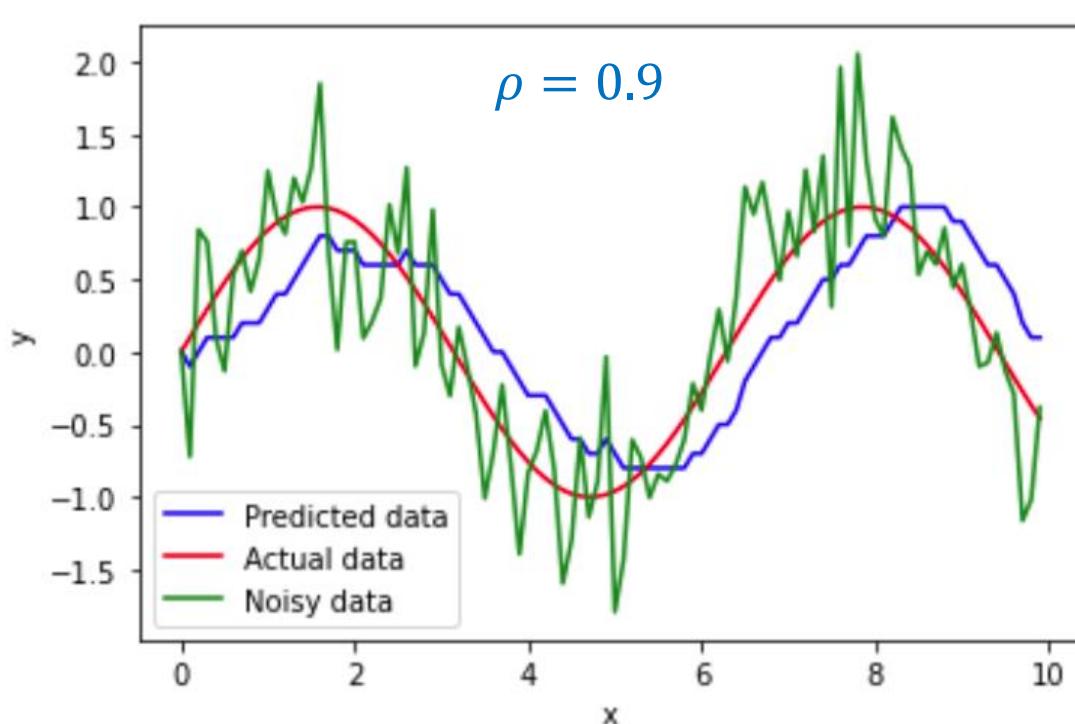
Exponentially weighted averages



Optimization Algorithms

Exponentially weighted averages

$$V_t = \rho V_{t-1} + (1 - \rho)s_t$$



Back to Adaptive Learning Rate

How to apply to adagrad

Adagrad

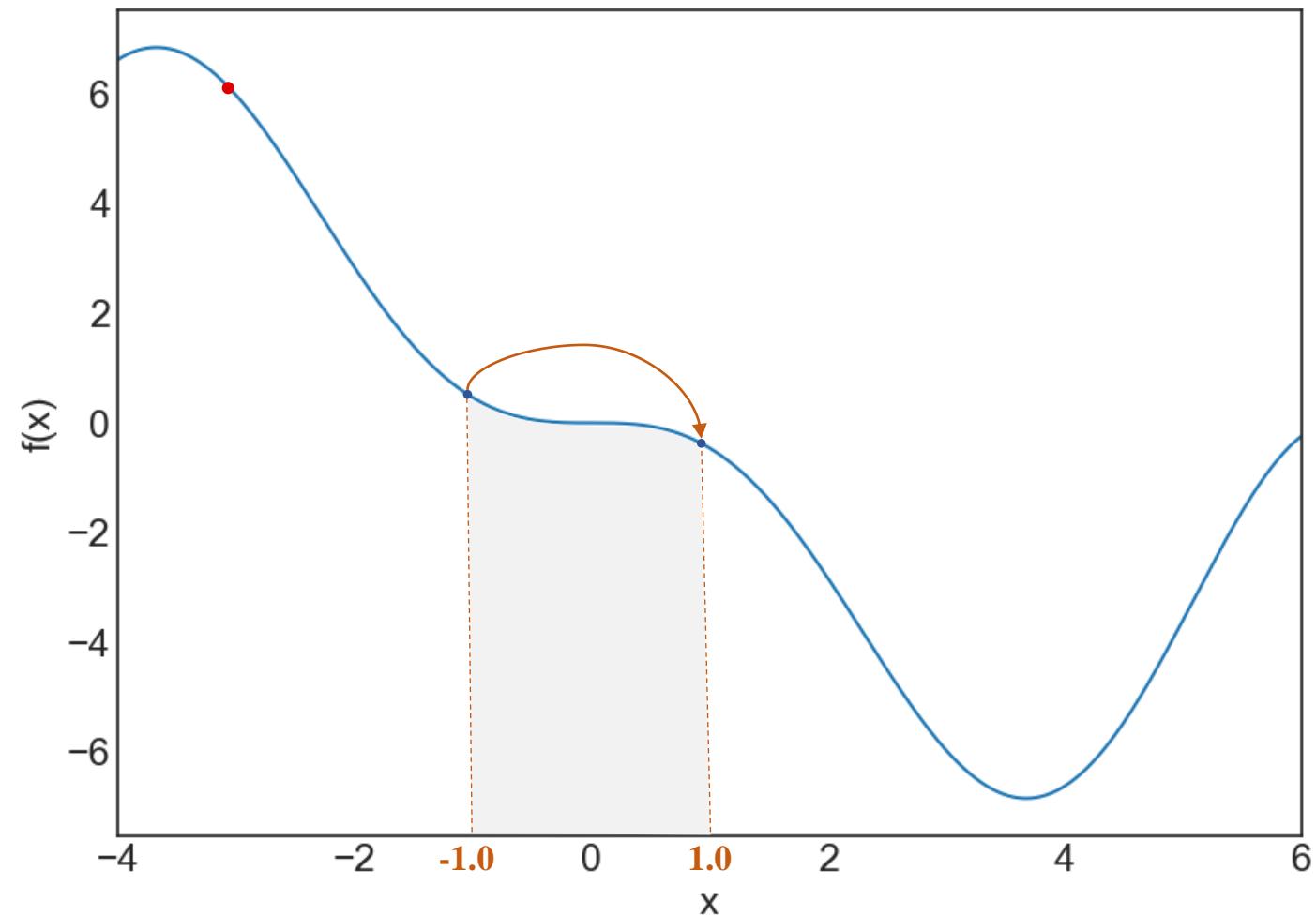
$$g_t = f'(x_{t-1})$$

$$s_t = s_{t-1} + g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

Exponentially weighted averages

$$V_t = \rho V_{t-1} + (1 - \rho) s_t$$



expected
summation

Back to Adaptive Learning Rate

How to apply to adagrad

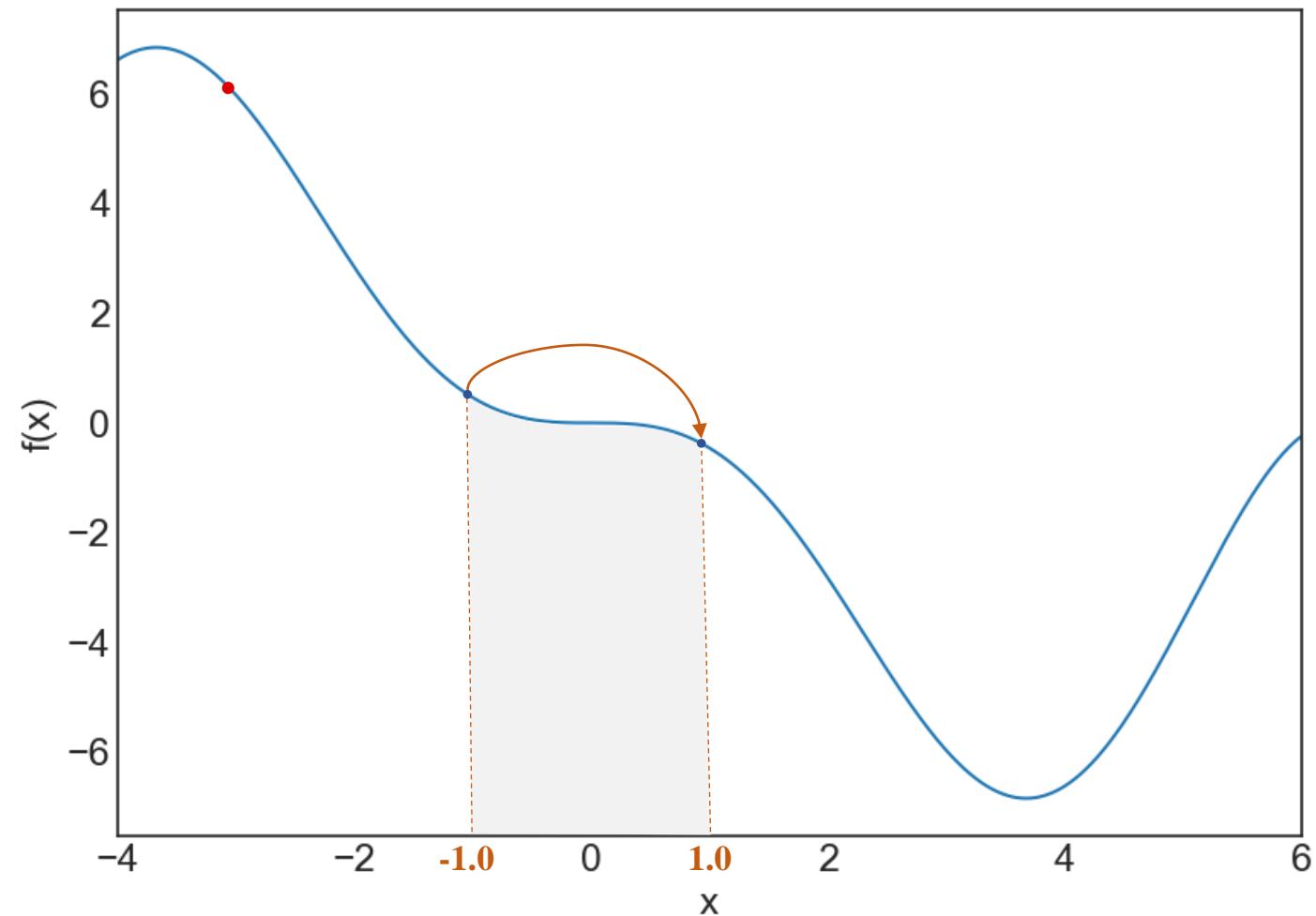
Adagrad RMSProp

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

expected summation



Back to Adaptive Learning Rate

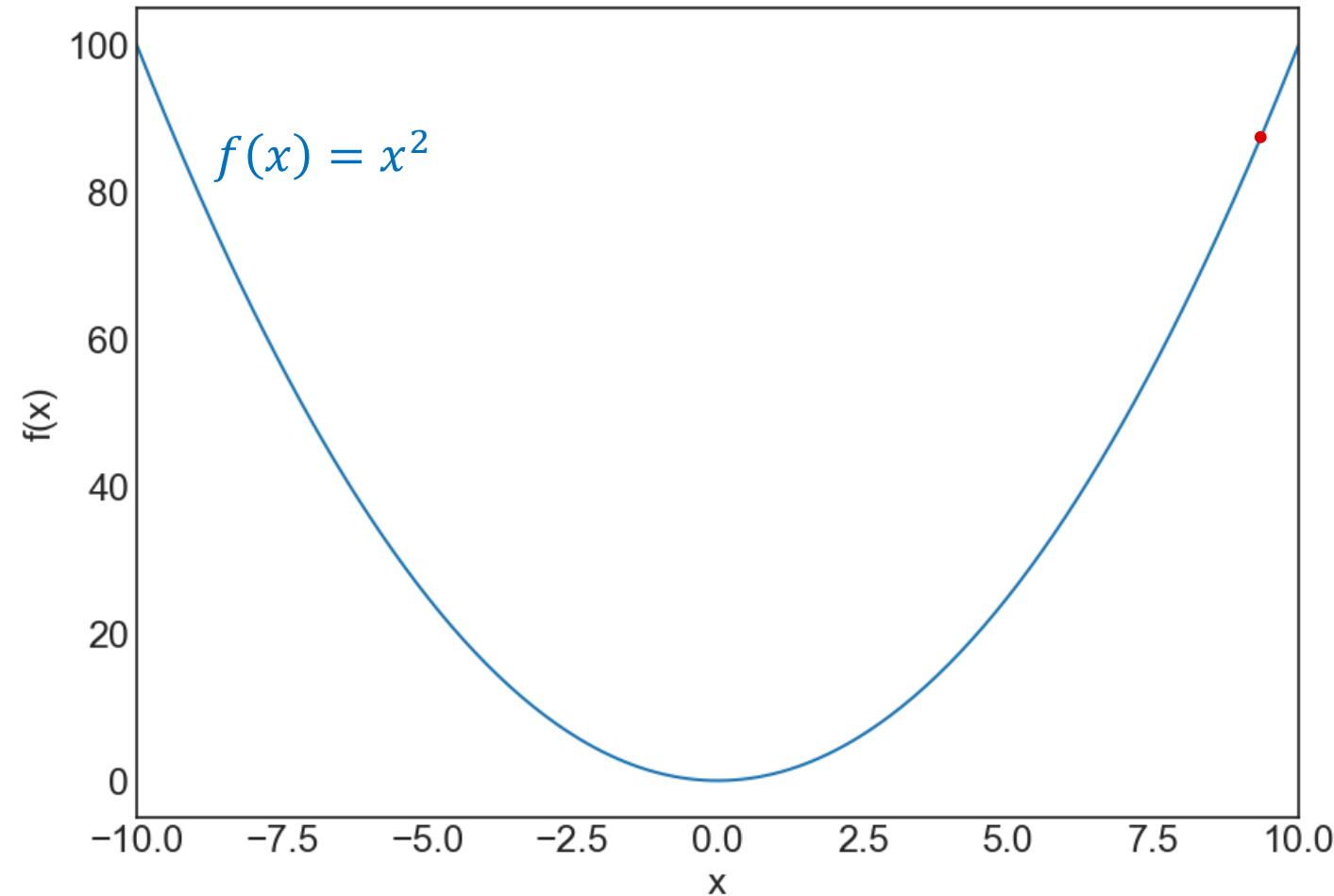
For square function

RMSProp

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_{t-1}$$



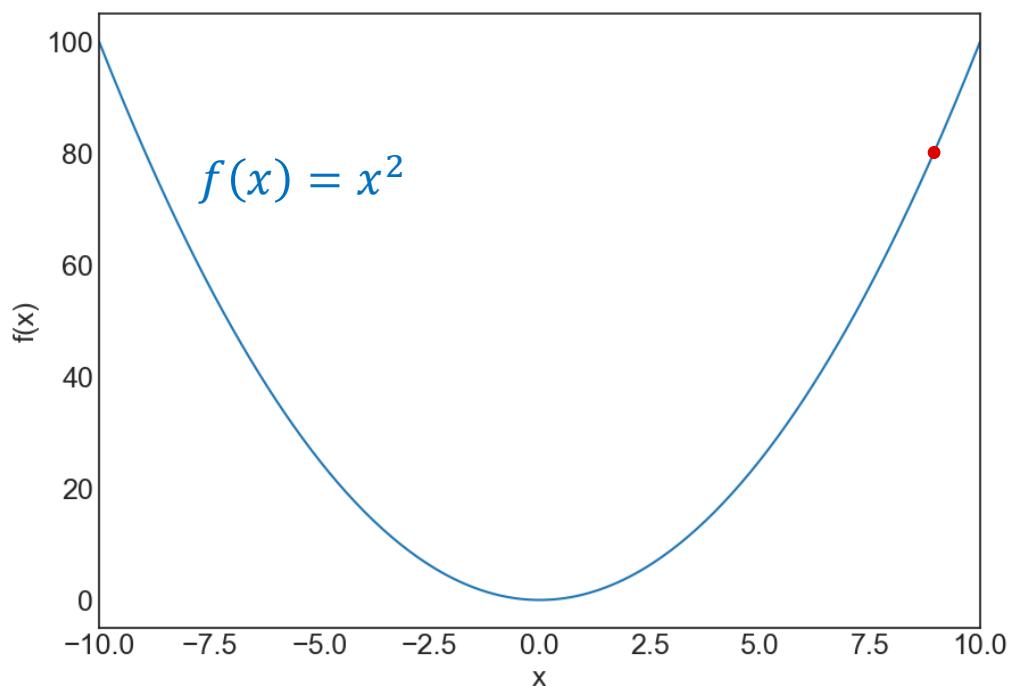
For square function

RMSProp

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$



$$x_0 = 9.0 \quad \eta = 0.1 \quad s_0 = 0.0$$

$$\rho = 0.9 \quad \epsilon = 10^{-7}$$

$$f'(x_0) = 18.0 \quad s_1 = 32.4$$

$$x_1 = 9.0 - \frac{0.1}{\sqrt{s_1 + \epsilon}} f'(x_0) = 8.6837$$

$$f'(x_1) = 17.36 \quad s_2 = 59.3$$

$$x_2 = 8.458$$

$$f'(x_2) = 16.916 \quad s_3 = 82.0$$

$$x_2 = 8.27$$

Back to Adaptive Learning Rate

For square function

RMSProp

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

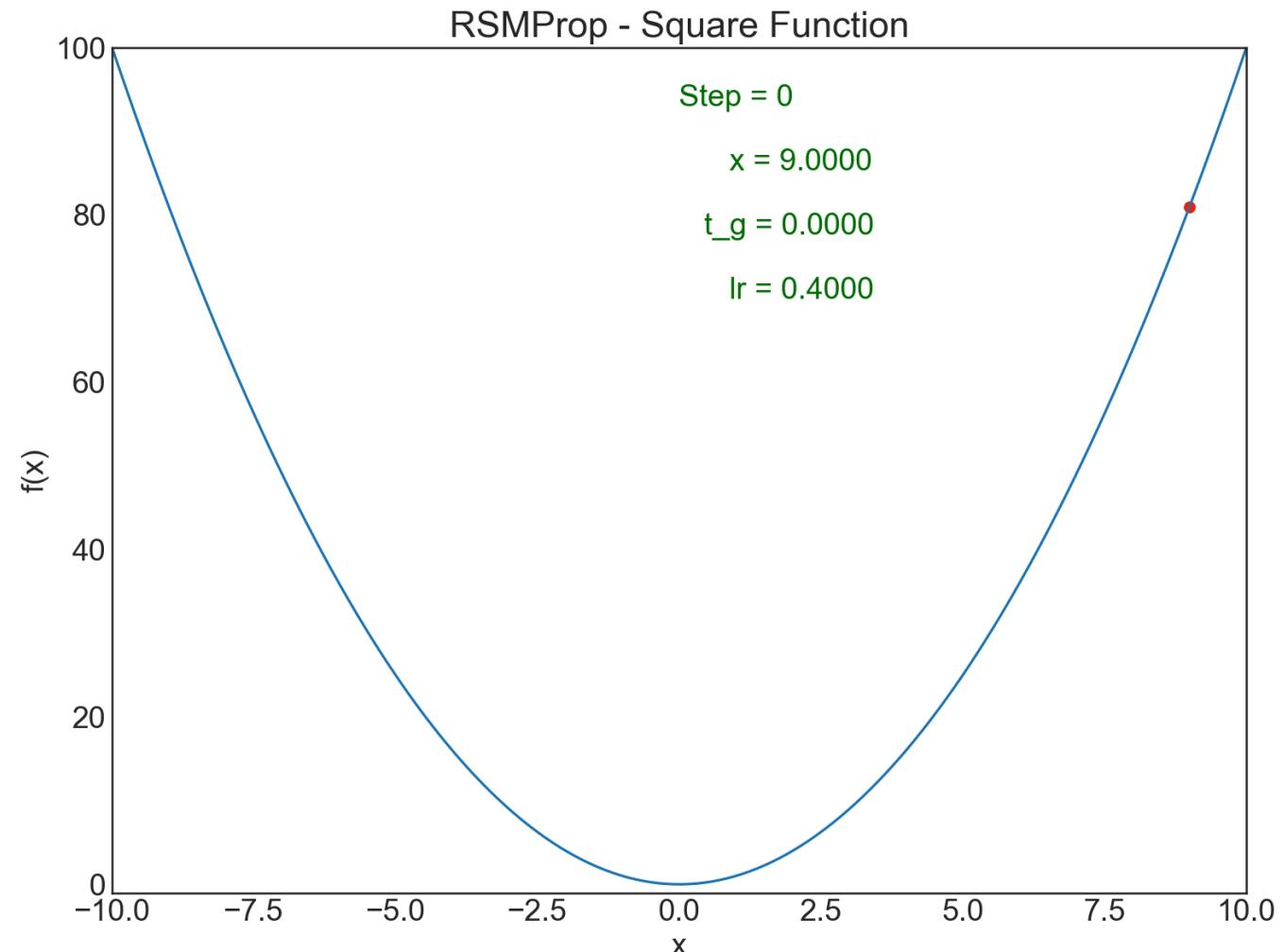
$$x_0 = 9.0$$

$$\eta = 0.4$$

$$s_0 = 0.0$$

$$\rho = 0.9$$

$$\epsilon = 10^{-7}$$



Back to Adaptive Learning Rate

For another function

RMSProp

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

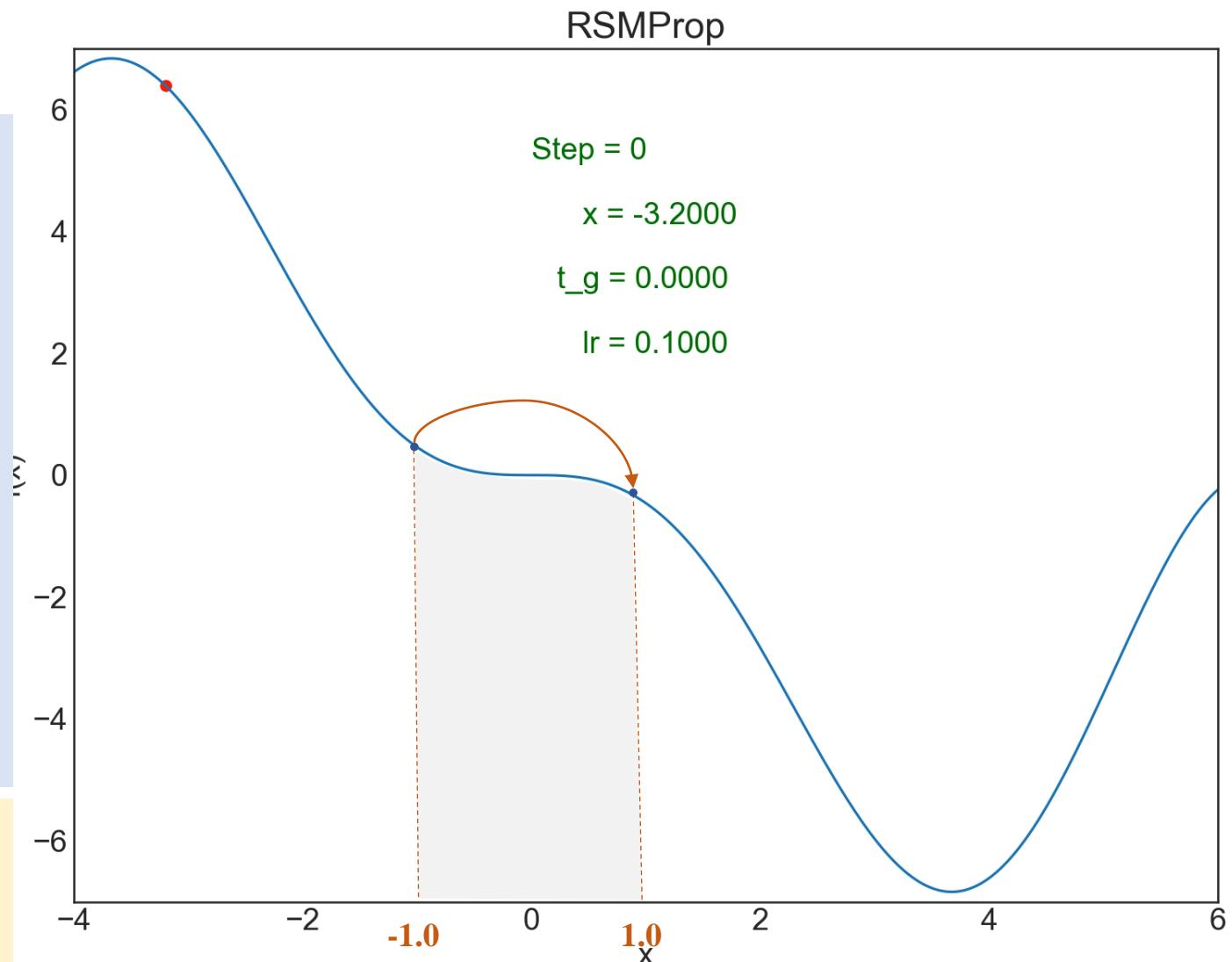
$$x_0 = -3.5$$

$$\eta = 0.1$$

$$s_0 = 0.0$$

$$\rho = 0.9$$

$$\epsilon = 10^{-7}$$



Adaptive Learning Rate

Generalization

RMSProp (One variable Function)

$$g_t = f'(x_{t-1})$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$

RMSProp (Multi-variable Function)

$$g_t = \nabla_{\theta} L$$

$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_{t,i}$$

Outline

- Quick Introduction
- Stochastic Gradient Descent
- Adagrad
- RMSProp
- Momentum
- Adam
- Case studies

Adaptive Learning Rate

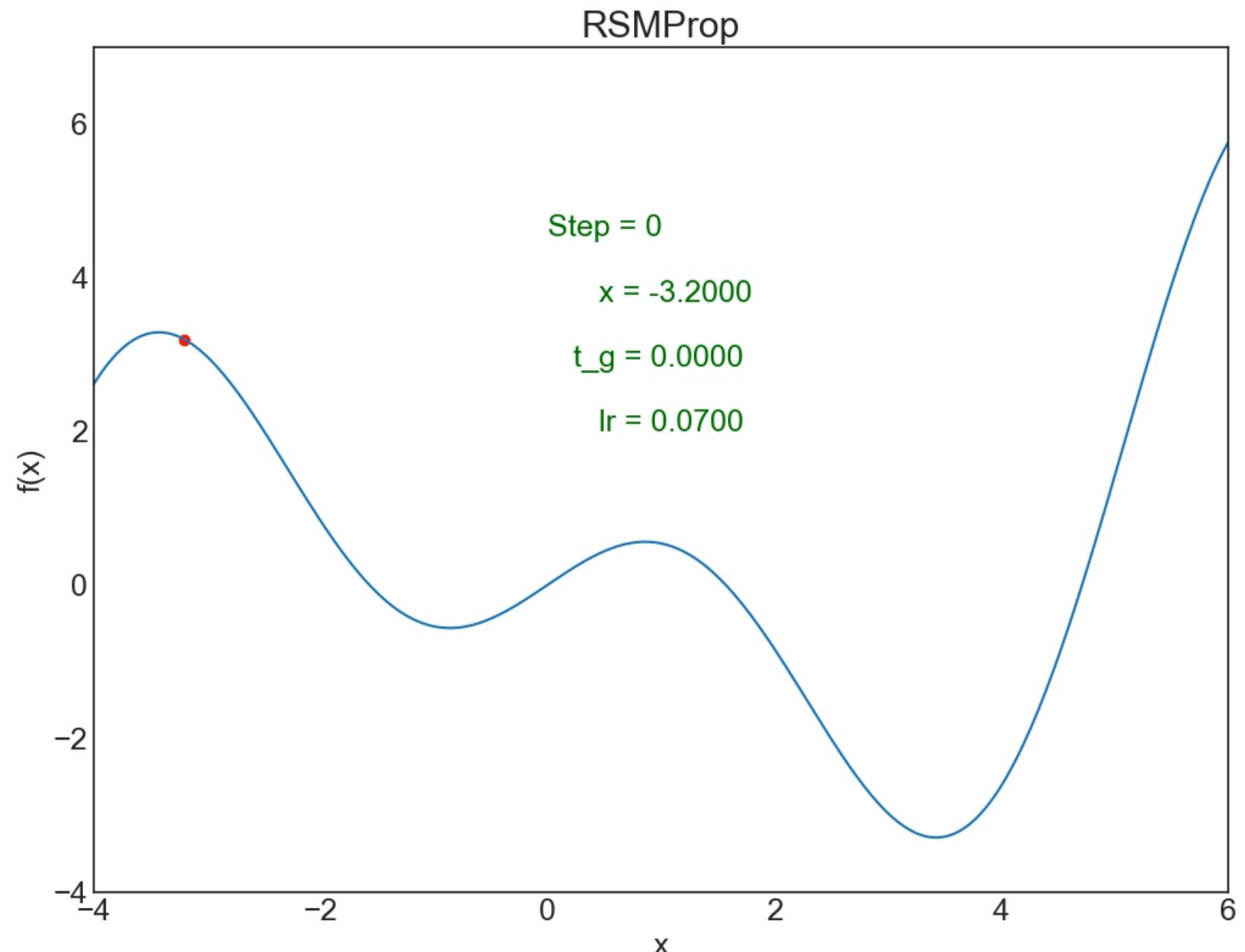
A common limitation so far

$$\theta_t = \theta_{t-1} - \eta \nabla_{\theta} L$$

$$g_t = \nabla_{\theta} L$$

$$s_t = s_{t-1} + g_t^2$$

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



Using Momentum

SGD

$$\theta_t = \theta_{t-1} - \eta \nabla_{\theta} L$$

SGD + Momentum

$$v_t = \rho v_{t-1} + (1 - \rho) \alpha \nabla_{\theta} L$$

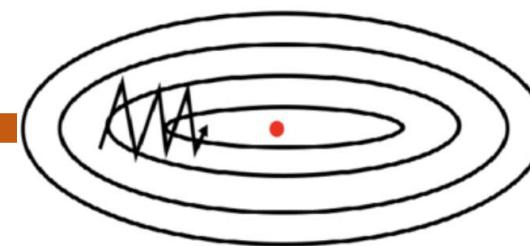
$$\theta_t = \theta_{t-1} - v_t$$



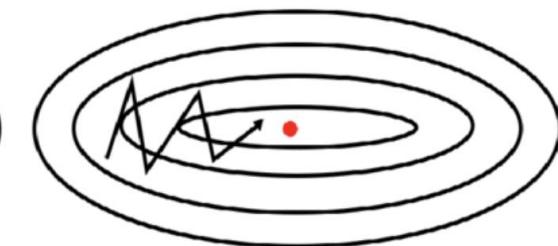
$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$

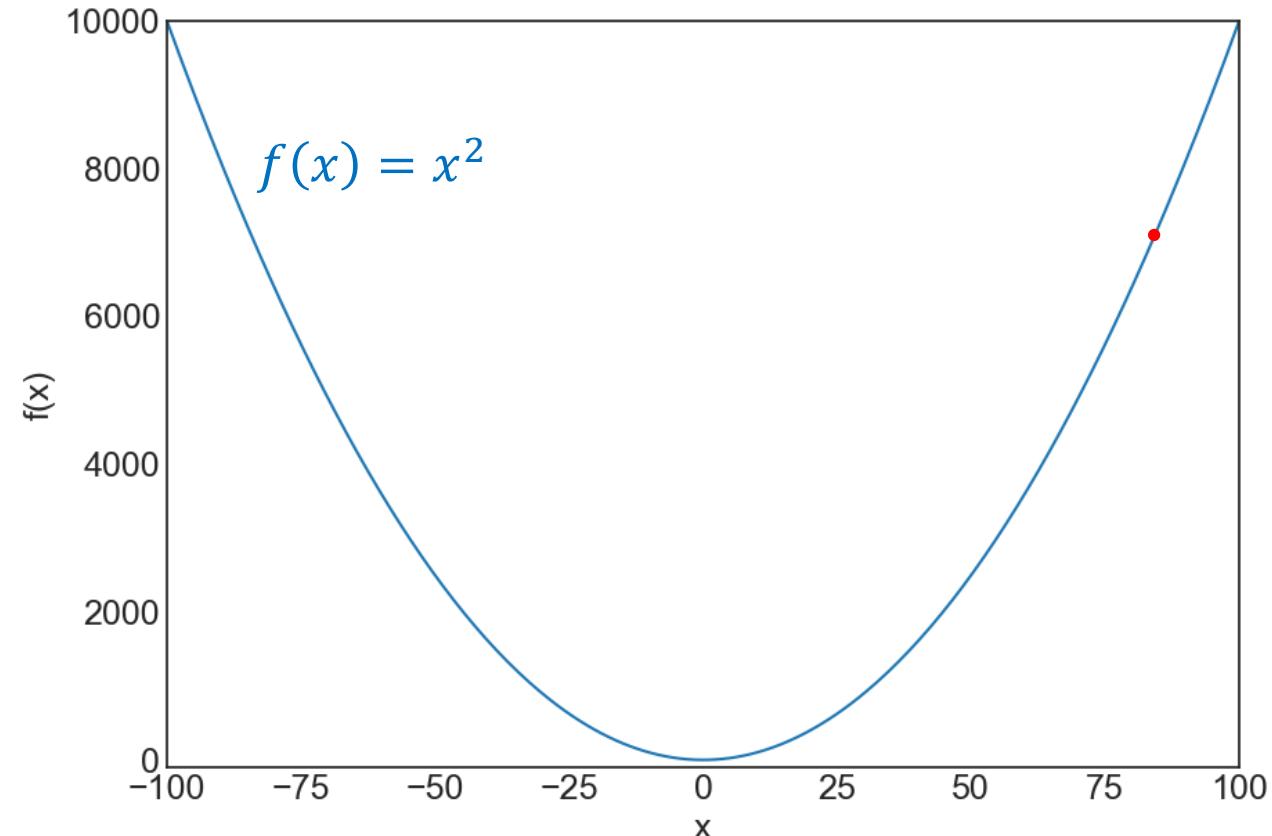
SGD without momentum



SGD with momentum



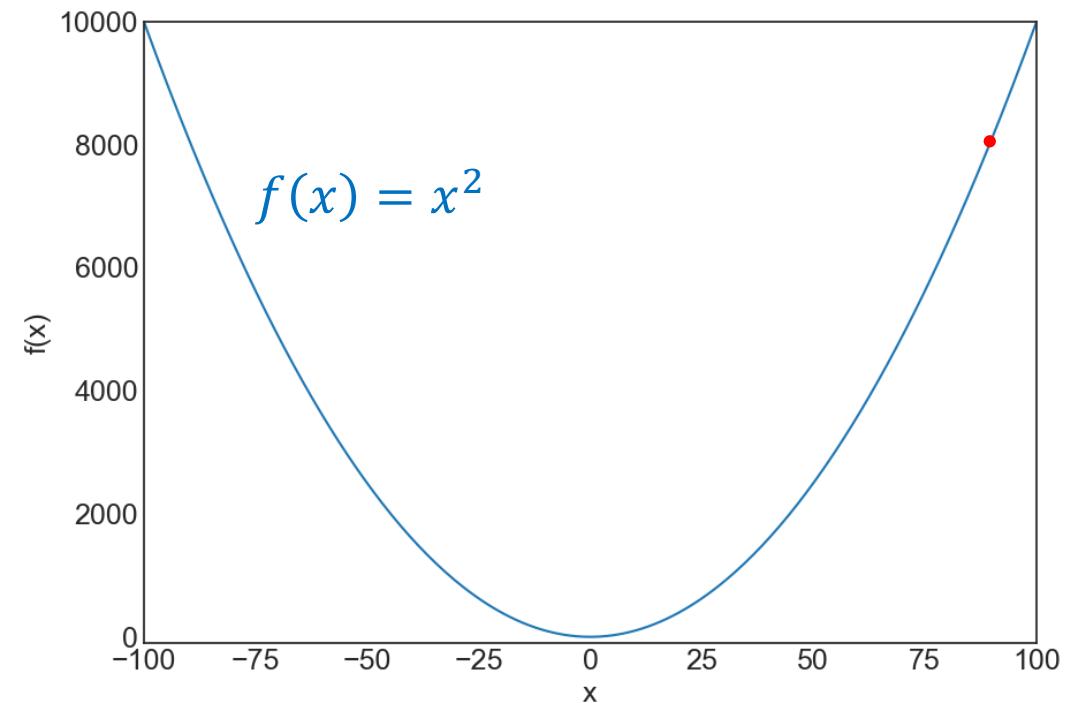
Genevieve B. Orr



SGD + Momentum

$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$



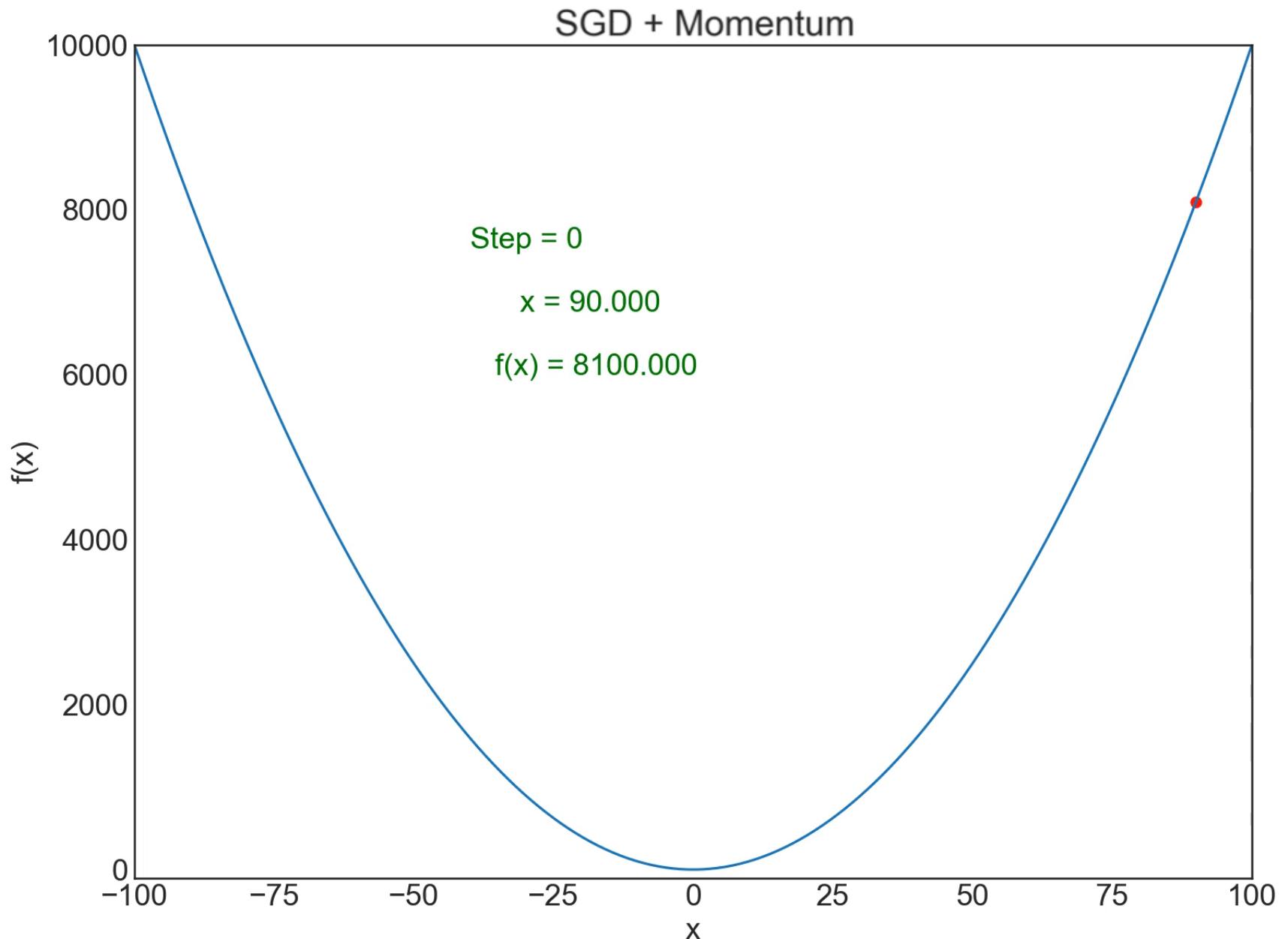
$x_0 = 90.0$	$m = 0.9$
$\eta = 0.1$	$v_0 = 0.0$
$f'(x_0) = 180.0$	$v_1 = -18.0$
$x_1 = x_0 + v_1 = 72.0$	
$f'(x_1) = 144.0$	$v_2 = -30.59$
$x_2 = x_1 + v_2 = 41.4$	
$f'(x_2) = 82.8$	$v_3 = -35.82$
$x_3 = x_2 + v_3 = 5.58$	
$f'(x_3) = 11.16$	$v_4 = -33.354$
$x_4 = x_3 + v_4 = -27.77$	
$f'(x_4) = -55.54$	$v_5 = -24.46$
$x_5 = x_4 + v_5 = -52.23$	

SGD + Momentum

$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$

$$\begin{array}{ll} x_0 = 90.0 & m = 0.0 \\ \eta = 0.03 & v_0 = 0.0 \end{array}$$

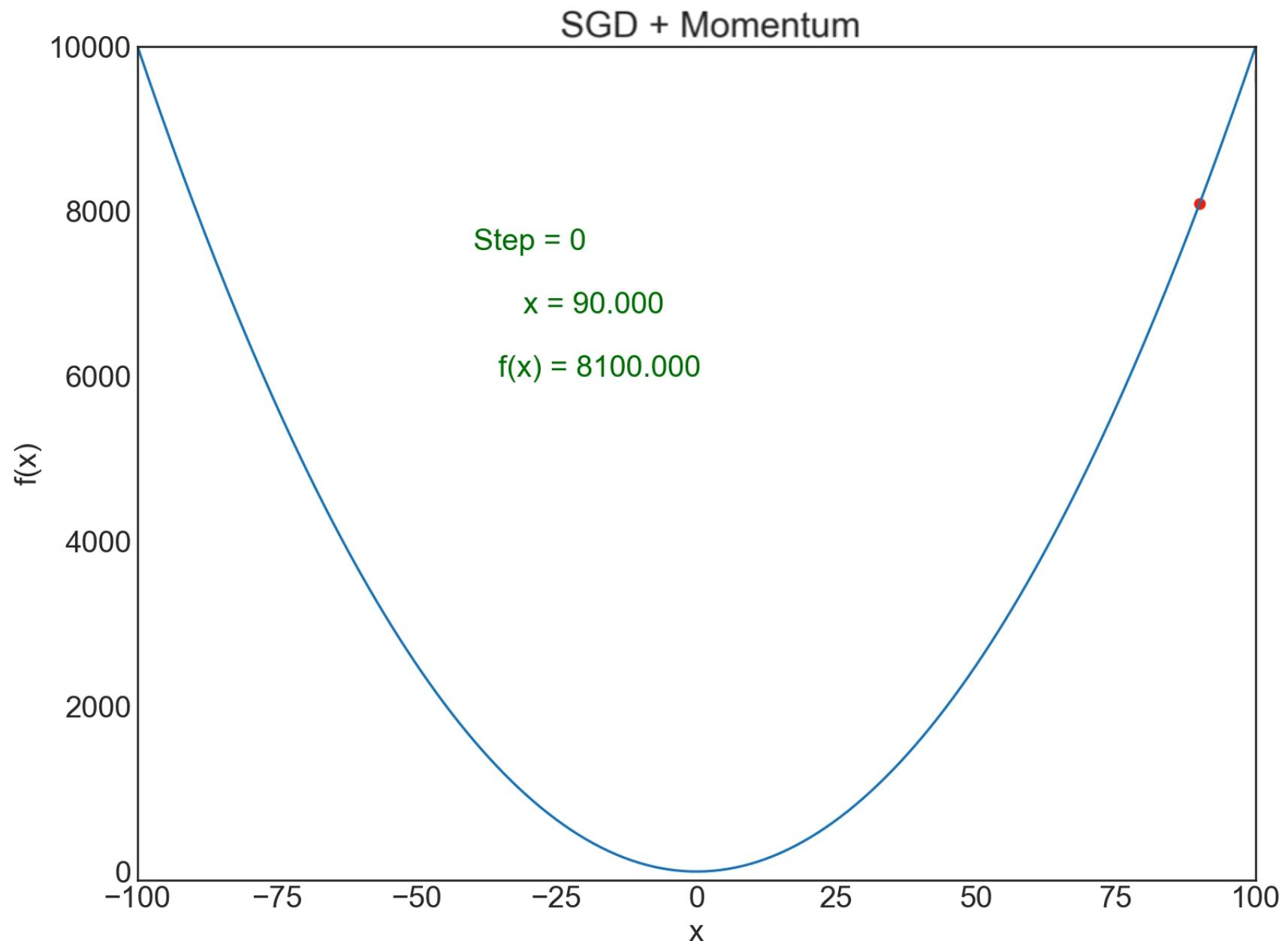


SGD + Momentum

$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

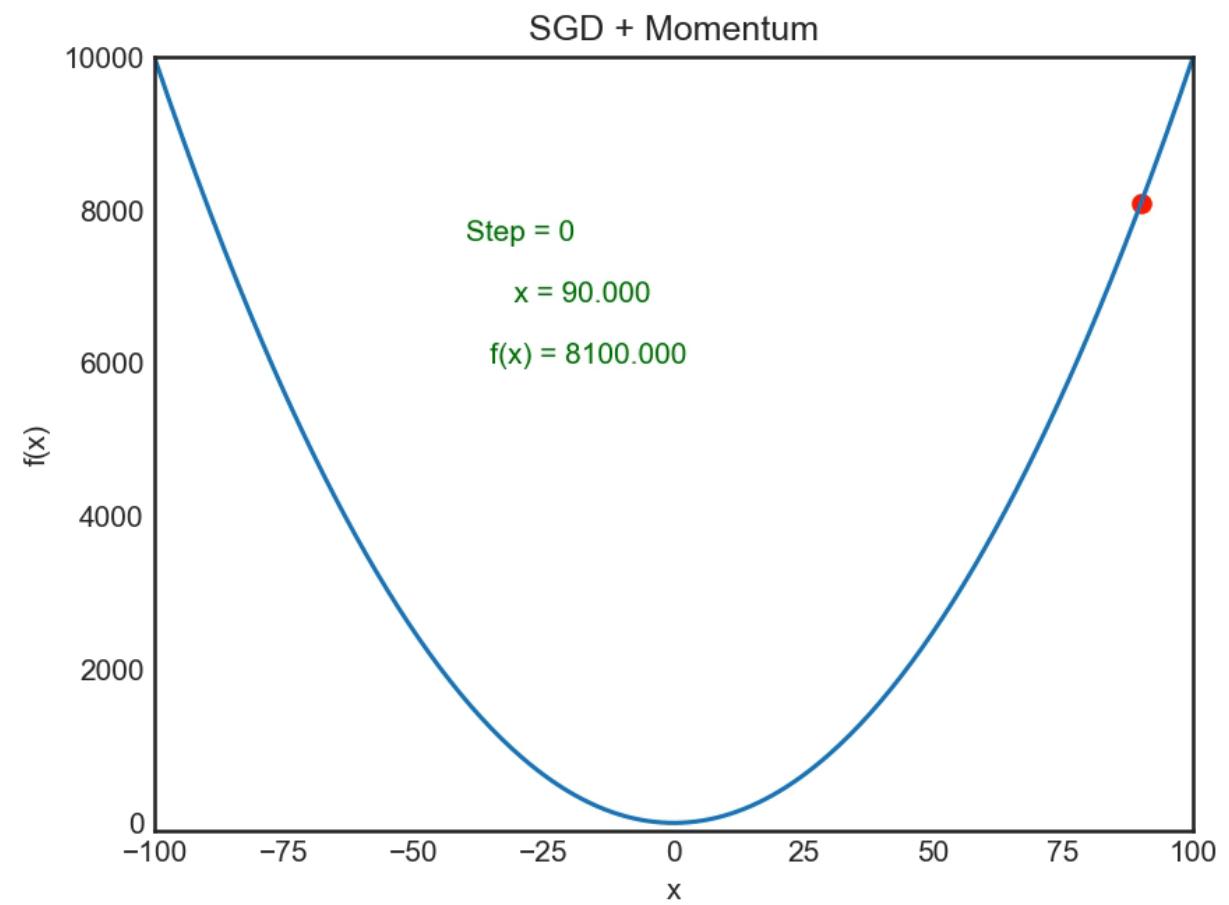
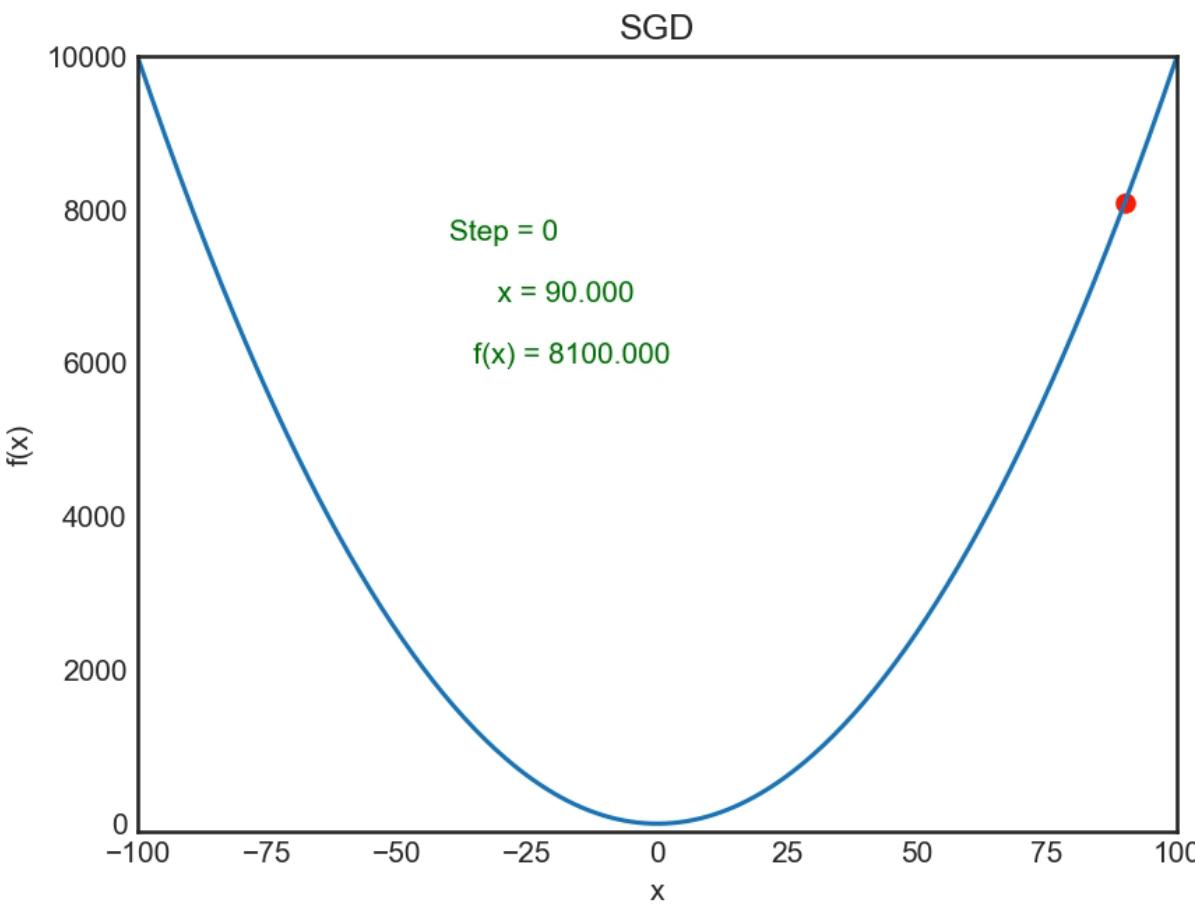
$$\theta_t = \theta_{t-1} + v_t$$

$$\begin{array}{ll} x_0 = 90.0 & m = 0.9 \\ \eta = 0.01 & v_0 = 0.0 \end{array}$$



SGD + Momentum

Mimic a ball rolling

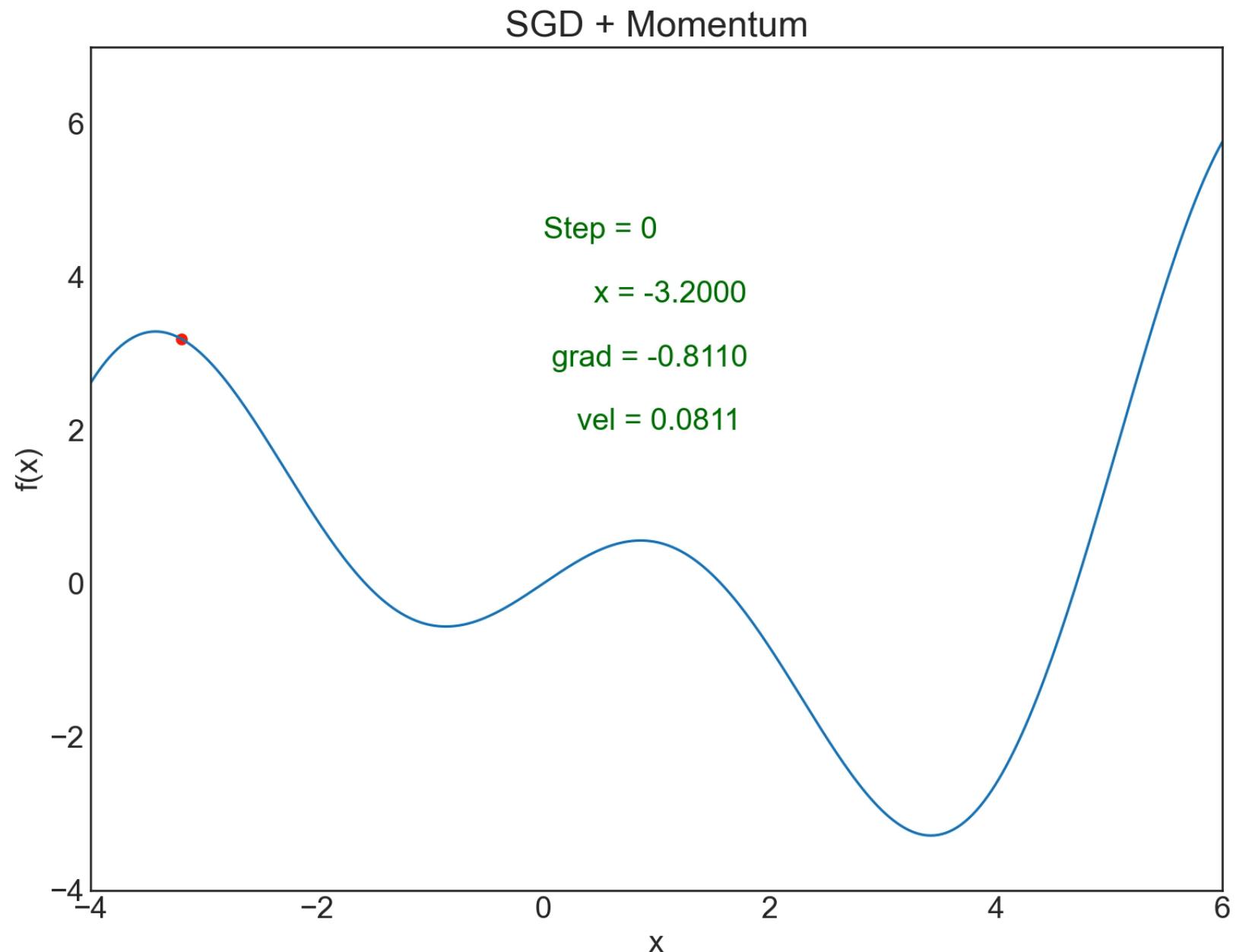


SGD + Momentum

$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$

$$\begin{aligned}x_0 &= -3.2 & m &= 0.9 \\ \eta &= 0.1 & v_0 &= 0.0\end{aligned}$$



What about RMSProp+Momentum

SGD

$$\theta_t = \theta_{t-1} - \eta \nabla_{\theta} L$$

SGD + Momentum

$$v_t = m v_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$

$$v_t = \rho v_{t-1} - (1 - \rho) \alpha \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + v_t$$

RMSProp

$$g_t = \nabla_{\theta} L$$
$$s_t = \rho s_{t-1} + (1 - \rho) g_t^2$$
$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{s_t + \epsilon}} g_t$$



What about RMSProp+Momentum

SGD + Momentum

$$m_t = \rho m_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + m_t$$

$$m_t = \rho m_{t-1} - (1 - \rho) \alpha \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + m_t$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} - \eta m_t$$

idea equivalence

RMSProp

$$\begin{aligned} g_t &= \nabla_{\theta} L \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\ \theta_t &= \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \epsilon}} g_t \end{aligned}$$

Simpler version of Adam

$$g_t = \nabla_{\theta} L$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \epsilon}} m_t$$

What about RMSProp+Momentum

SGD + Momentum

$$m_t = \rho m_{t-1} - \eta \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + m_t$$

$$m_t = \rho m_{t-1} - (1 - \rho) \alpha \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} + m_t$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla_{\theta} L$$

$$\theta_t = \theta_{t-1} - \eta m_t$$

idea equivalence

Further reading

RMSProp

$$g_t = \nabla_{\theta} L$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$
$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \epsilon}} g_t$$

Adam

$$g_t = \nabla_{\theta} L$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\frac{v_t}{1 - \beta_2^t} + \epsilon}} \times \frac{m_t}{1 - \beta_1^t}$$

Outline

- Quick Introduction
- Stochastic Gradient Descent
- Adagrad
- RMSProp
- Momentum
- Adam
- Case studies

Image Data

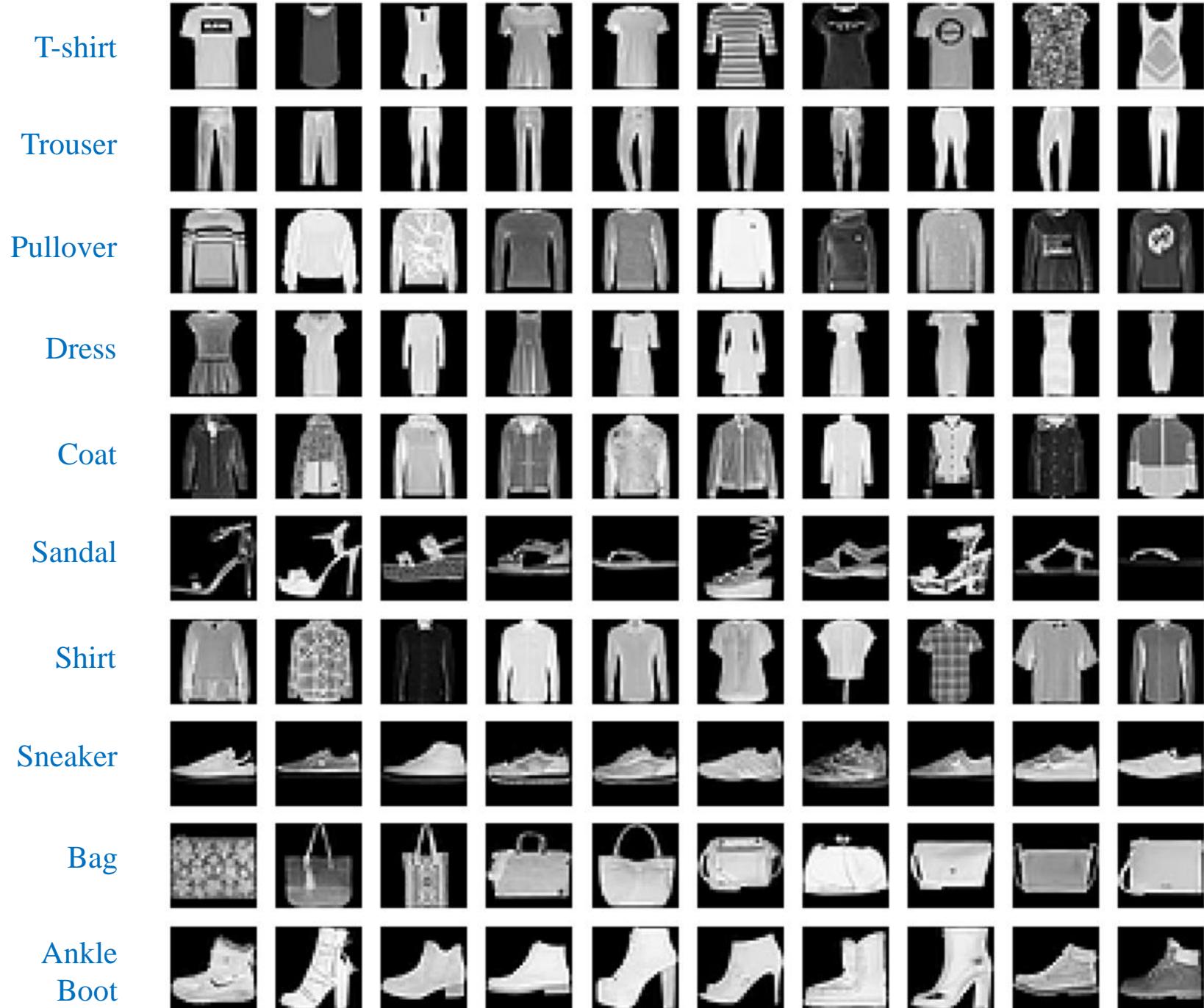
Fashion-MNIST dataset

Grayscale images

Resolution=28x28

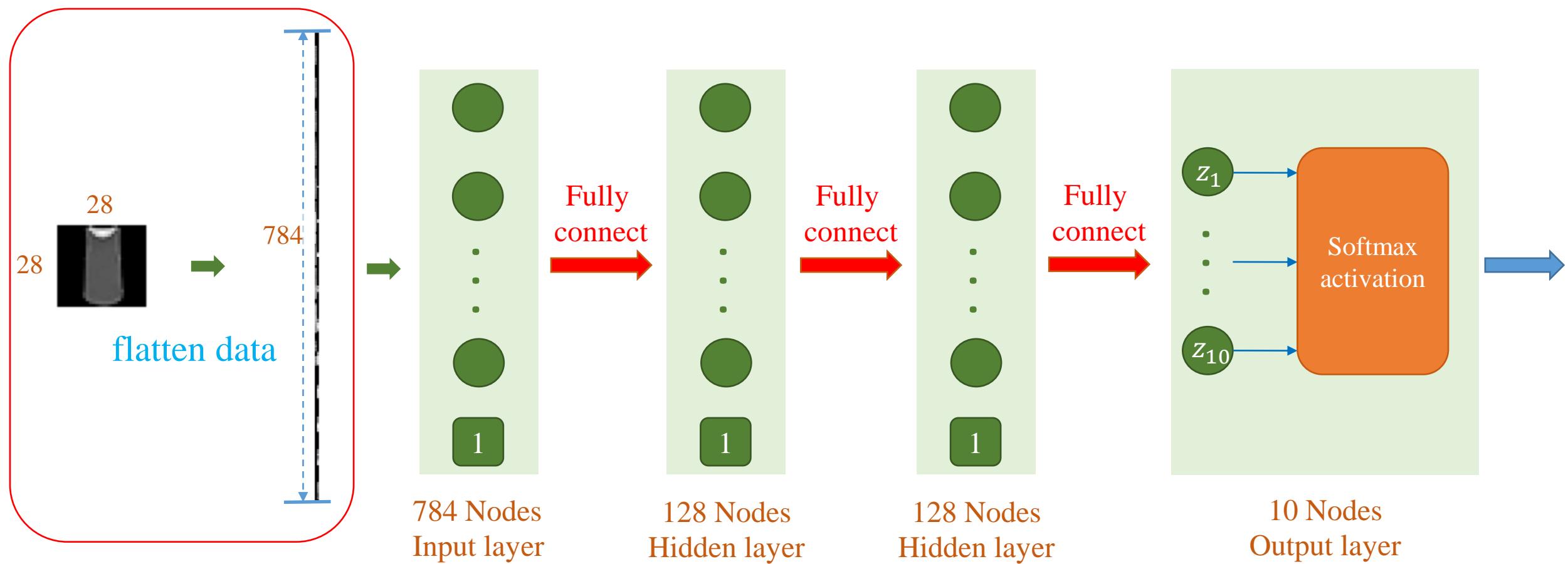
Training set: 60000 samples

Testing set: 10000 samples



Case study 1

Fashion-MNIST



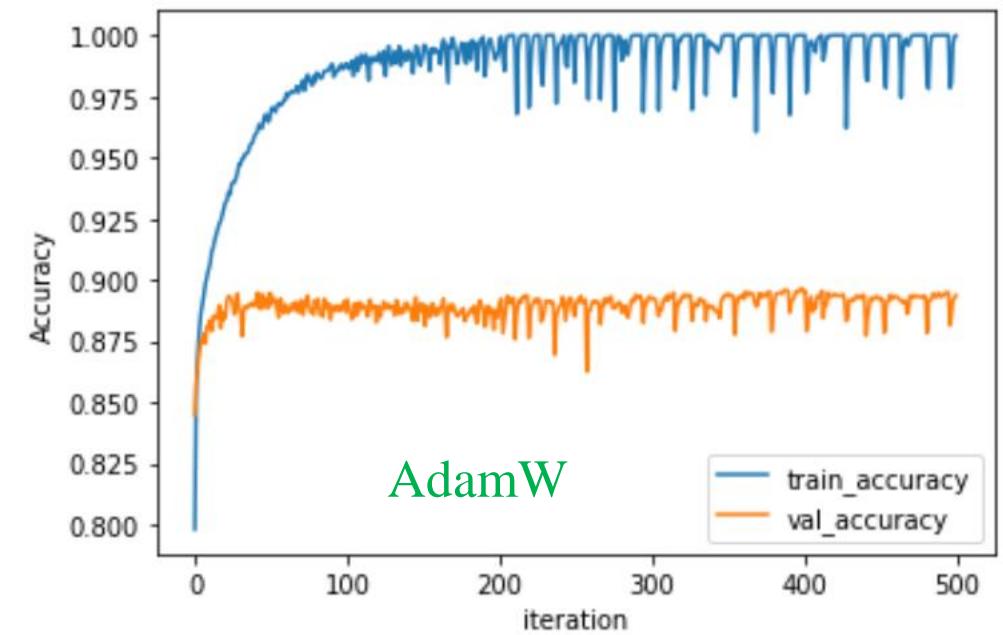
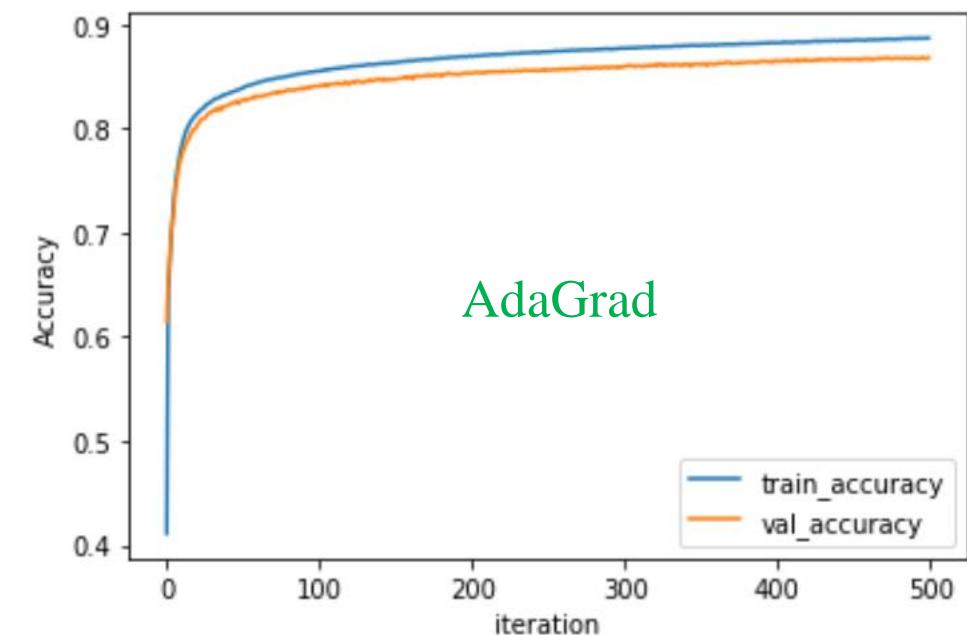
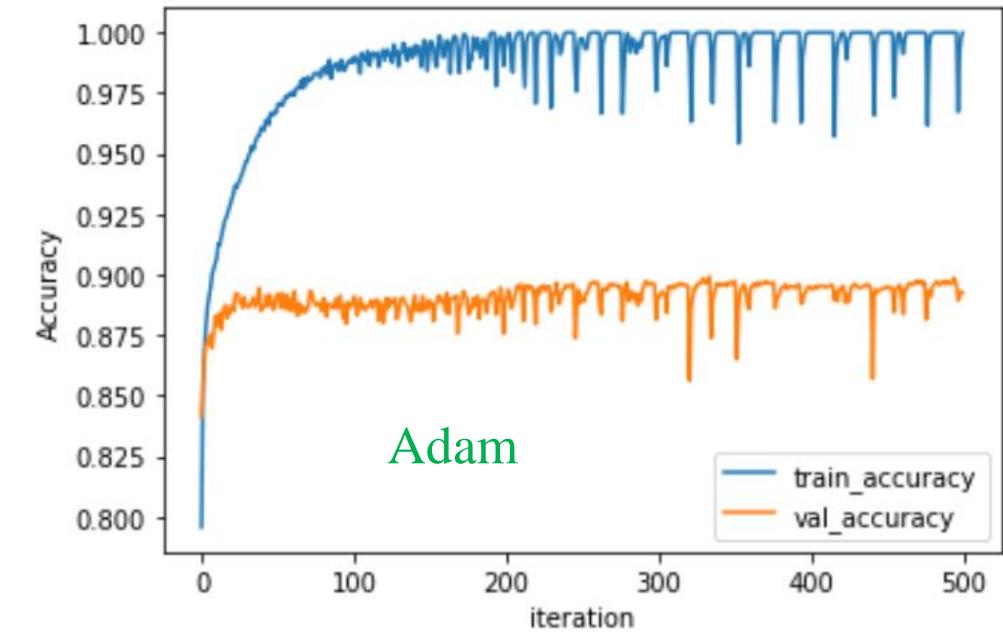
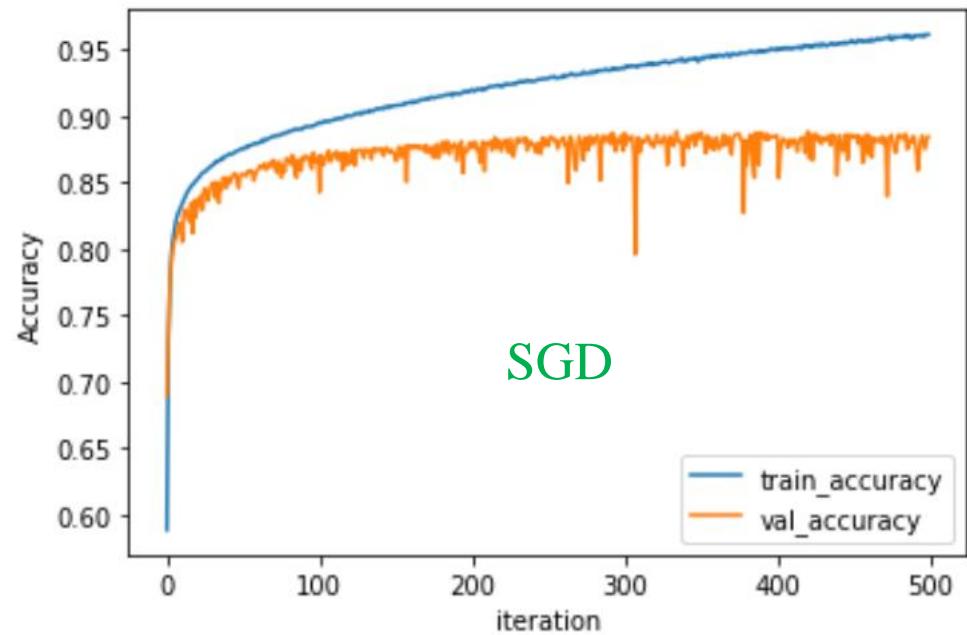


Image Data

Cifar-10 dataset

Color images

Resolution=32x32

Training set: 50000 samples

Testing set: 10000 samples

airplane



automobile



bird



cat



deer



dog



frog



horse



ship

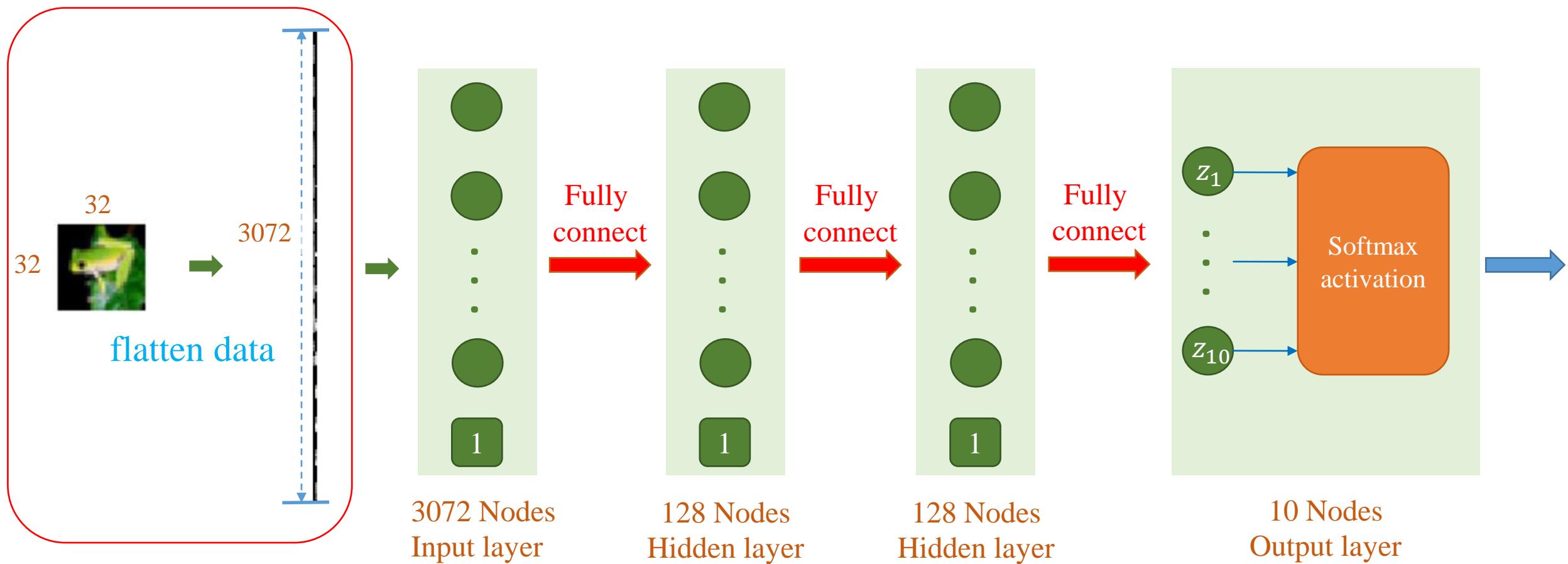


truck



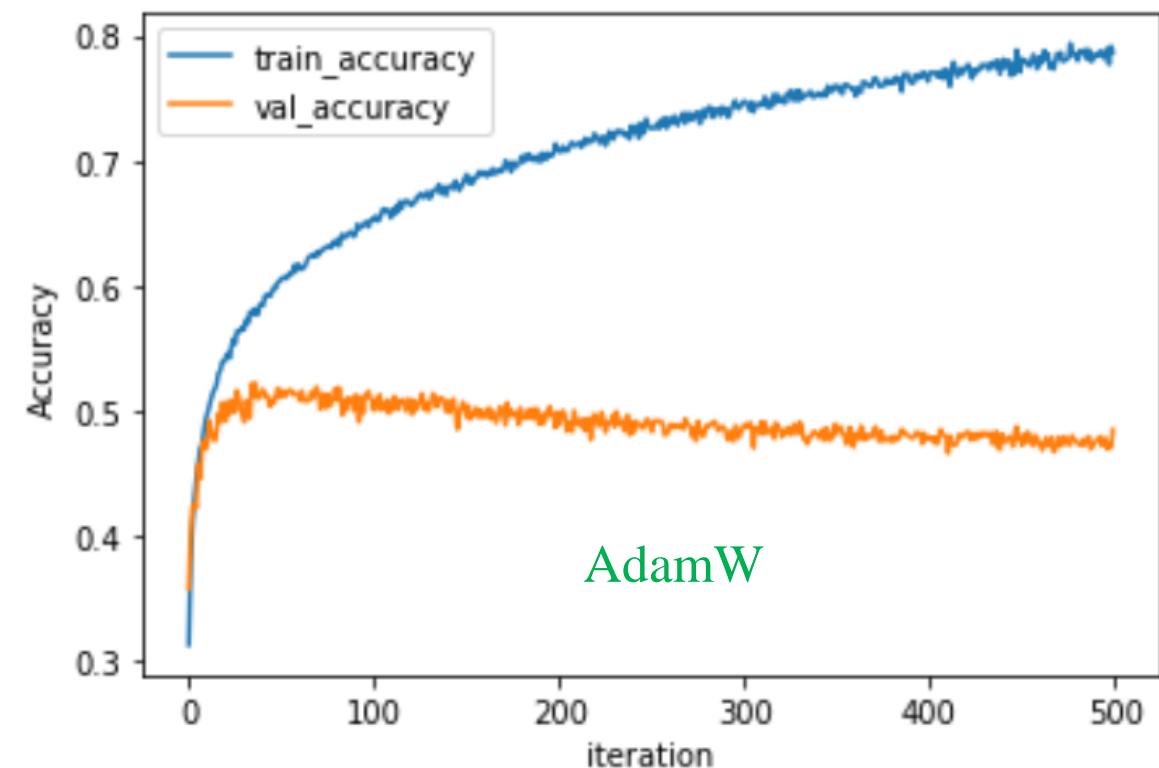
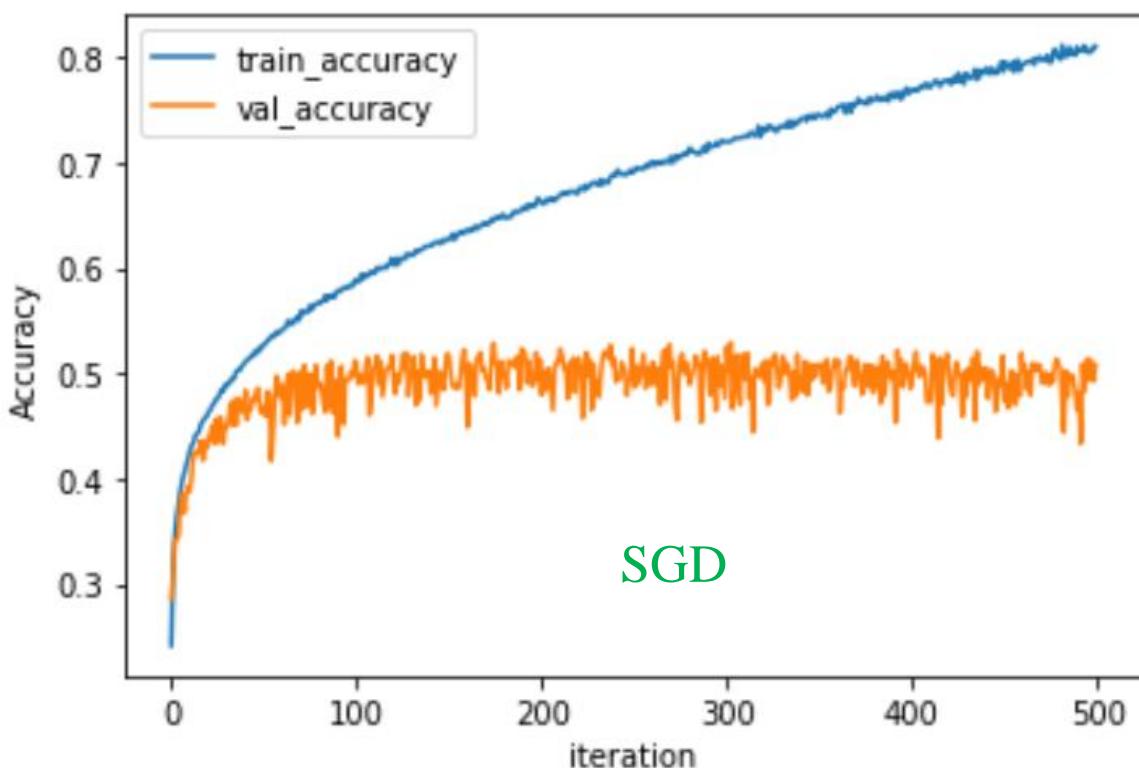
Case study 2

Cifar10-MNIST



Case Studies

Results



Further Reading

Bias Correction

<http://manjeetdahiya.com/posts/exponential-weighted-average/>

<https://stats.stackexchange.com/questions/232741/why-is-it-important-to-include-a-bias-correction-term-for-the-adam-optimizer-for>

Adam

<https://towardsdatascience.com/complete-guide-to-adam-optimization-1e5f29532c3d>

<https://optimization.cbe.cornell.edu/index.php?title=Adam>

