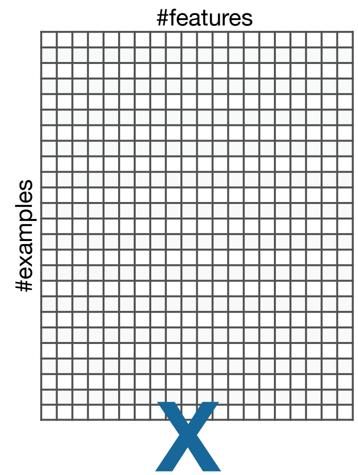
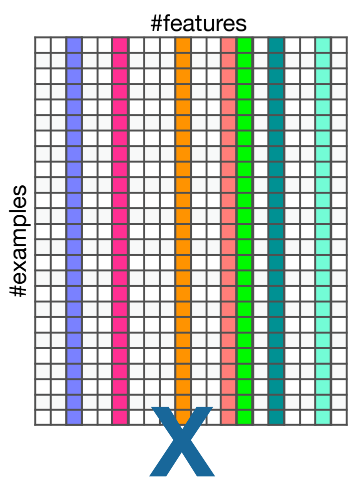
COS234:
Introduction To Machine Learning

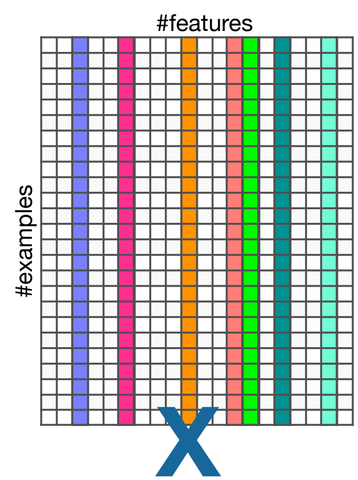
Prof. Yoram Singer

PRINCETON UNIVERSITY

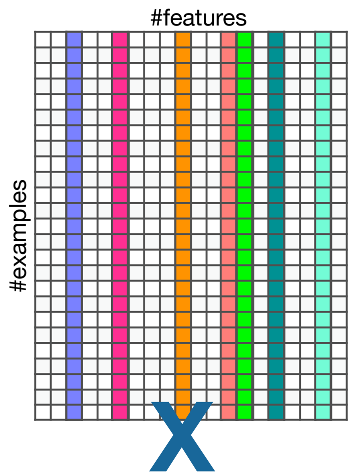
Topic: Gradient-Based Learning - Part I



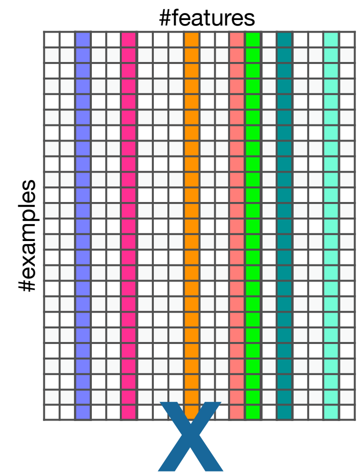




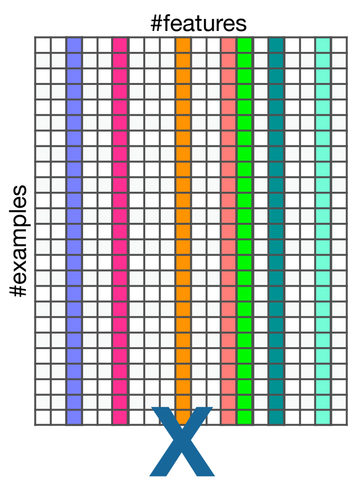
- Sequential: a column at a time



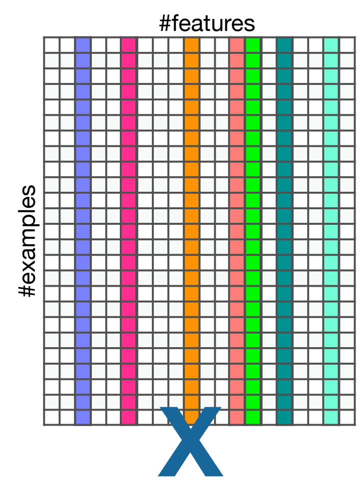
- Sequential: a column at a time
- Oblivious to “similar” examples



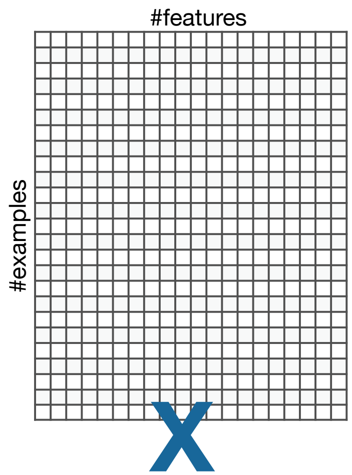
- Sequential: a column at a time
- Oblivious to “similar” examples
- Difficult to parallelize

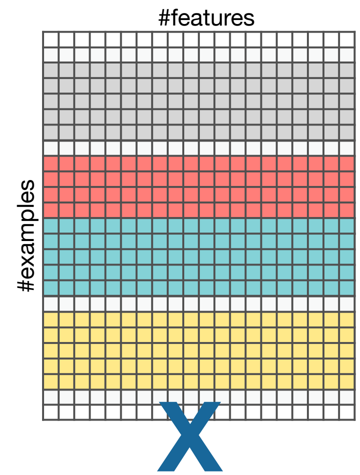


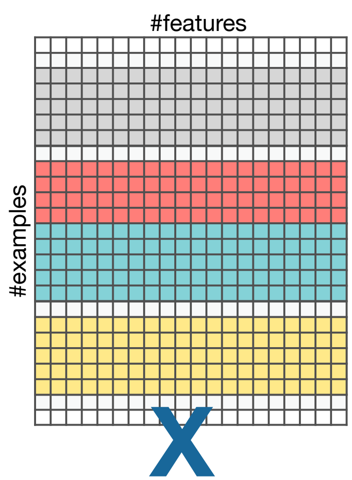
- Sequential: a column at a time
- Oblivious to “similar” examples
- Difficult to parallelize
- Requires dedicated update per loss



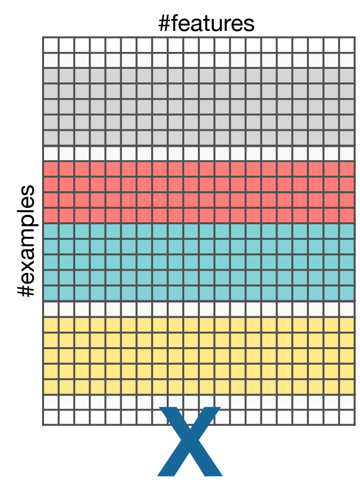
- Sequential: a column at a time
- Oblivious to “similar” examples
- Difficult to parallelize
- Requires dedicated update per loss
- Fails to work in non-linear settings



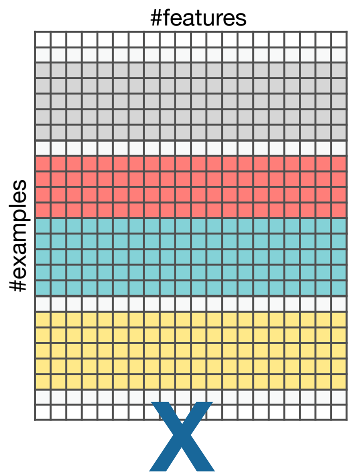




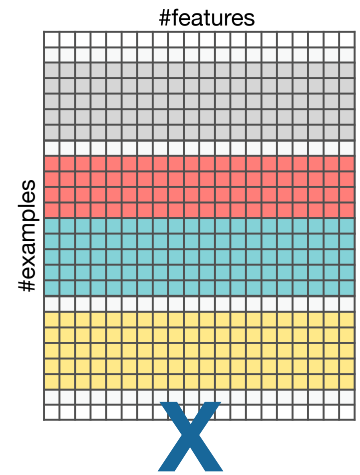
- Pick a small subset of rows of X to use



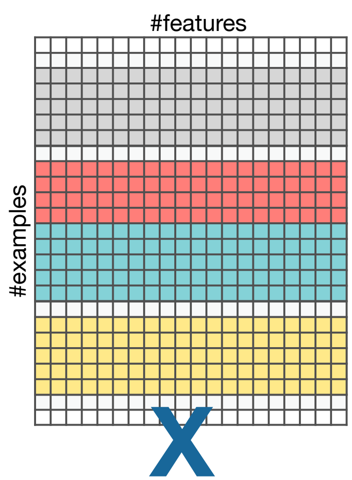
- Pick a small subset of rows of X to use
- Subset not necessarily consecutive rows



- Pick a small subset of rows of X to use
- Subset not necessarily consecutive rows
- Subset used to update all weights



- Pick a small subset of rows of X to use
- Subset not necessarily consecutive rows
- Subset used to update all weights
- Update "local" to subset selected



- Pick a small subset of rows of X to use
- Subset not necessarily consecutive rows
- Subset used to update all weights
- Update "local" to subset selected
- Albeit local, effect is "global"

- a^* minimum point of f:

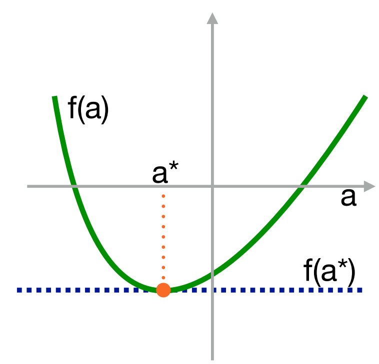
  $f(a^\star) \leq f(b)$ for all $b \neq a^\star$

- Derivate of f(a) at a^* is zero

  $$\left.\frac{df}{da}\right|_{a=a^\star} \equiv f'(a^\star) = 0$$

- Often no closed form solution for:

  $$\frac{df}{da} = 0$$



we can still make use of f' …

- $a^\star$ minimum point of f:

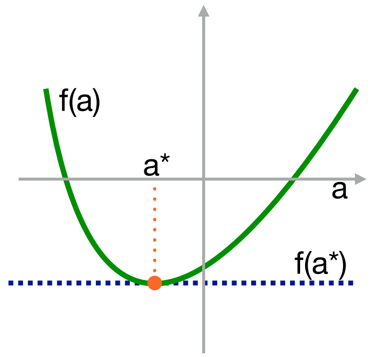
  $f(a^\star) \leq f(b)$ for all $b \neq a^\star$

- Derivate of f(a) at $a^\star$ is zero

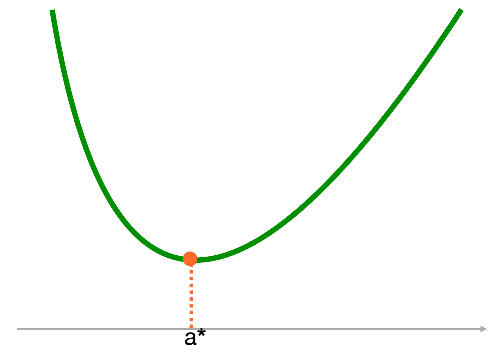
$$\frac{df}{da}\Big|_{a=a^\star} \equiv f'(a^\star) = 0$$

- Often no closed form solution for:

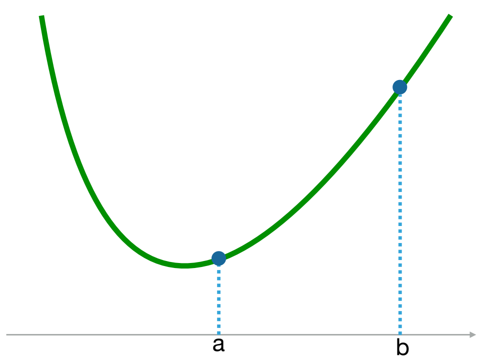
$$\frac{df}{da} = 0$$



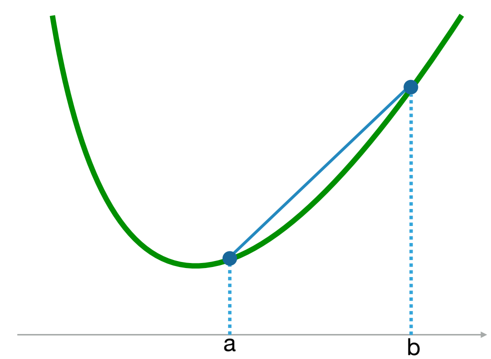
(Strict) Convexity



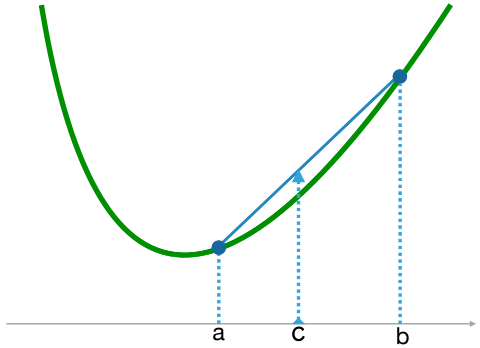
(Strict) Convexity



(Strict) Convexity

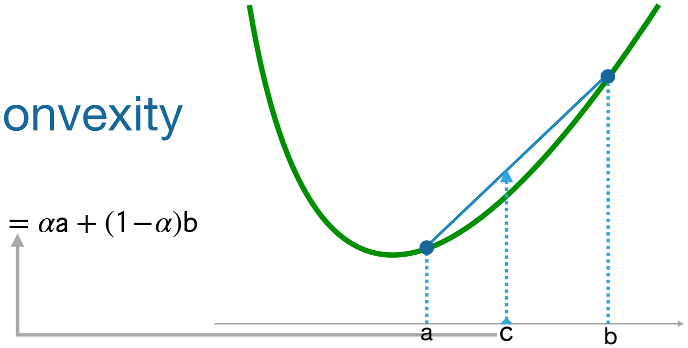


(Strict) Convexity



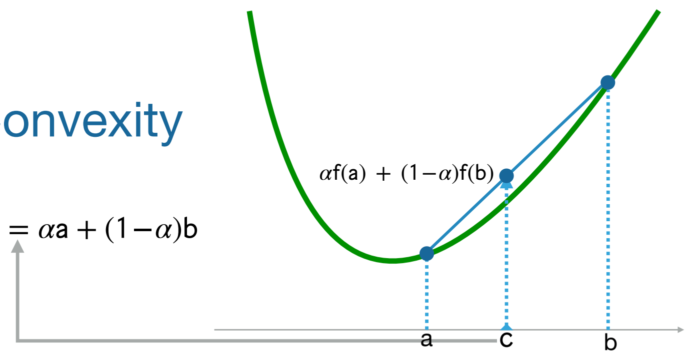
(Strict) Convexity

$$c = \alpha a + (1 - \alpha) b$$

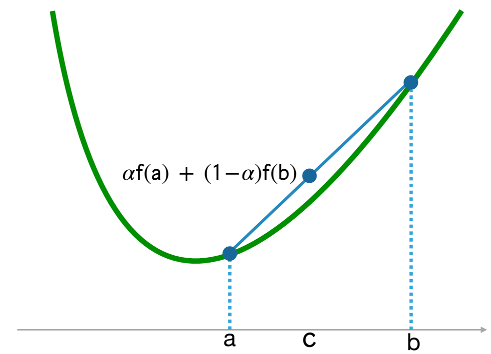


(Strict) Convexity

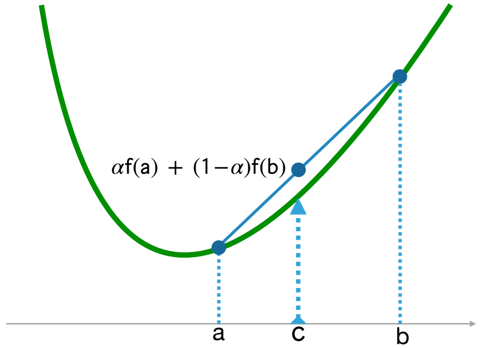
$$c = \alpha a + (1 - \alpha) b$$



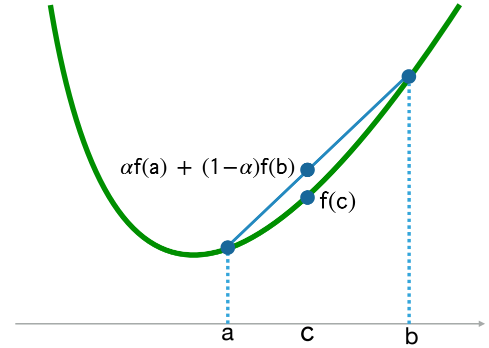
(Strict) Convexity



(Strict) Convexity

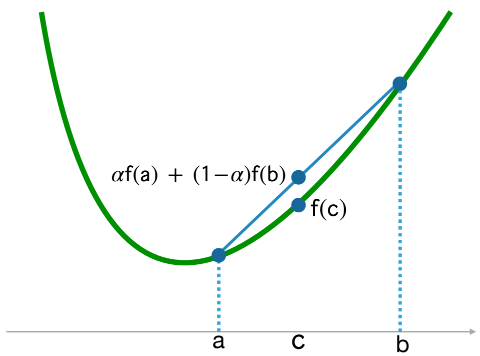


(Strict) Convexity



$$\alpha \in (0,1): \ \mathrm{f}\big(\alpha \mathrm{a} + (1-\alpha)\mathrm{b}\big) < \alpha \mathrm{f(a)} + (1-\alpha)\mathrm{f(b)}$$

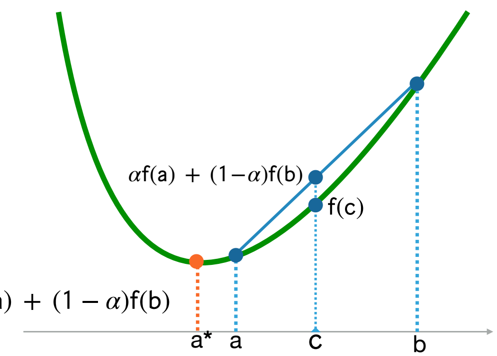
(Strict) Convexity

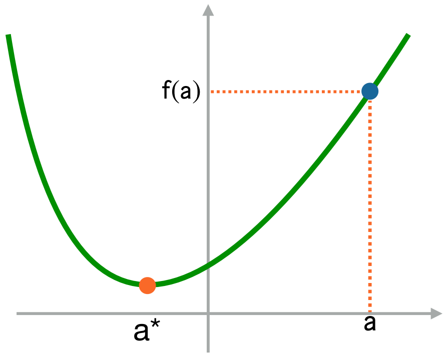


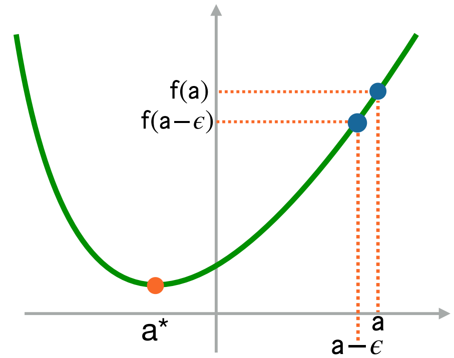
(Strict) Convexity

- a* is a unique minimum
- derivative is < 0 left to a*
- derivative is > 0 right to a*
- derivate of f(a) at a* is 0
- second derivative f''(a)>0

$$\alpha \in (0,1): \ \mathrm{f}\big(\alpha \mathrm{a} + (1-\alpha)\mathrm{b}\big) < \alpha \mathrm{f(a)} + (1-\alpha)\mathrm{f(b)}$$

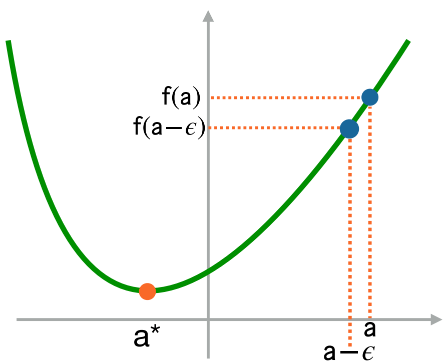






- if a to the right (a > a*) of a* then

$$f(a) > f(a-\epsilon)$$

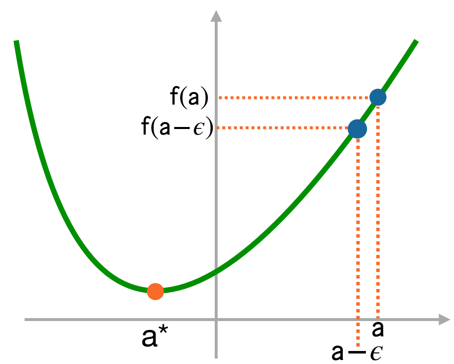


- if a to the right (a > a*) of a* then

$$f(a) > f(a-\epsilon)$$

- Therefore

$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$



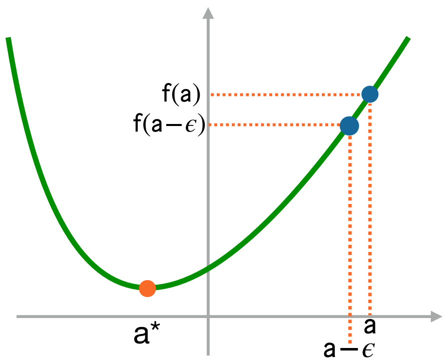
- if **a** to the right (**a** > a*) of a* then

$$f(a) > f(a-\epsilon)$$

- Therefore

$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$

- Taking $\epsilon \to 0$ we get $f'(a) > 0$



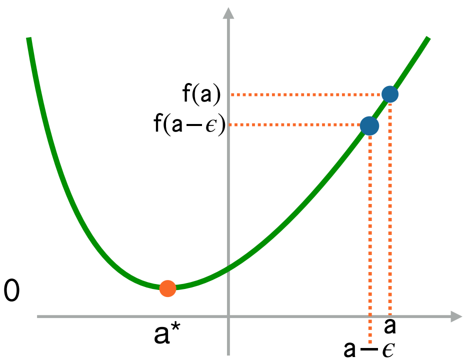
- if **a** to the right (**a** > a*) of a* then

$$f(a) > f(a-\epsilon)$$

- Therefore

$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$

- Taking $\epsilon \to 0$ we get $f'(a) > 0$
- Similarly for **a** < a* we get $f'(a) < 0$



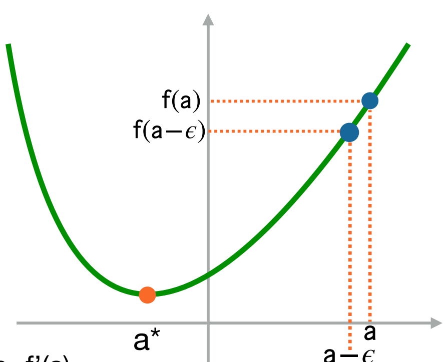
- if **a** to the right (**a** > a*) of a* then

$$f(a) > f(a-\epsilon)$$

- Therefore

$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$

- Taking $\epsilon \to 0$ we get $f'(a) > 0$
- Similarly for **a** < a* we get $f'(a) < 0$
- Get closer to a* by going in direction -f'(a)



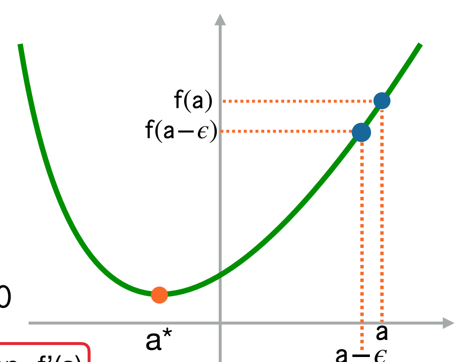
- if **a** to the right (**a** > a*) of a* then

$$f(a) > f(a-\epsilon)$$

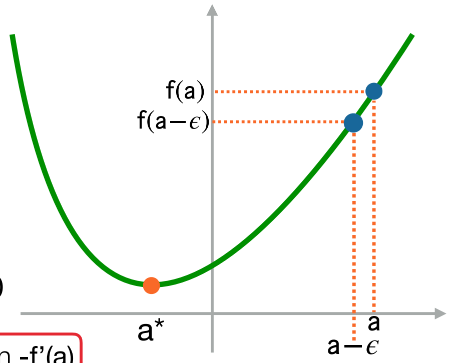
- Therefore

$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$

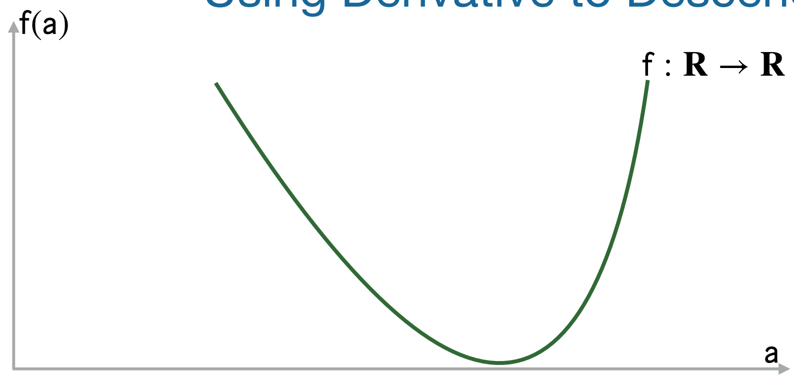
- Taking $\epsilon \to 0$ we get $f'(a) > 0$
- Similarly for **a** < a* we get $f'(a) < 0$
- Get closer to a* by going in direction -f'(a)



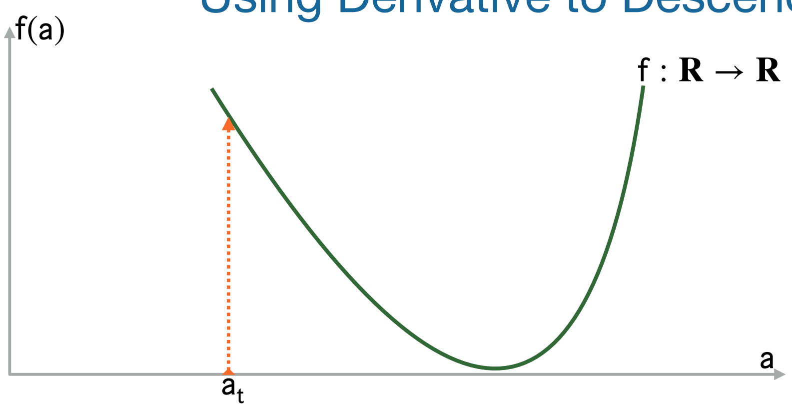
- if a to the right (a > a*) of a* then
$$f(a) > f(a-\epsilon)$$
- Therefore
$$\frac{f(a) - f(a-\epsilon)}{\epsilon} > 0$$
- Taking $\epsilon \rightarrow 0$ we get $f'(a) > 0$
- Similarly for a < a* we get $f'(a) < 0$
- $\boxed{\text{Get closer to a* by going in direction -f'(a)}}$



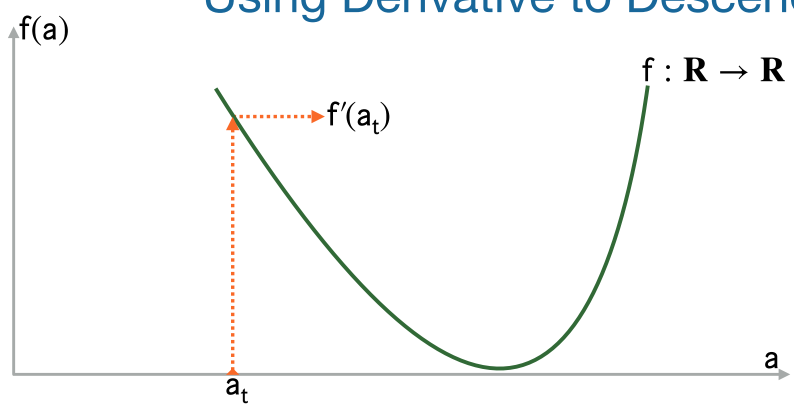
Using Derivative to Descend



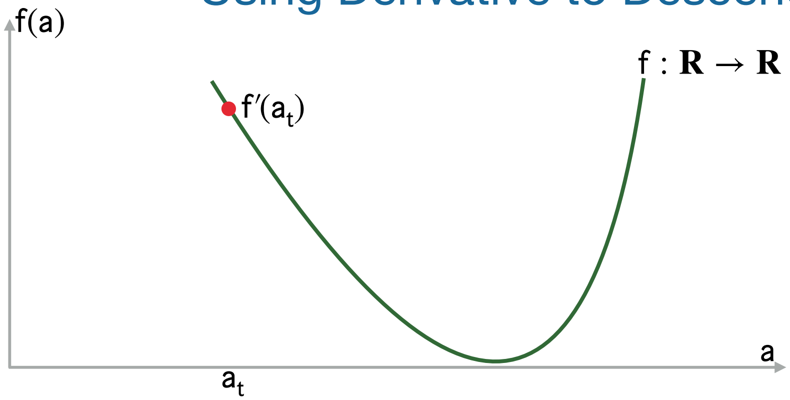
Using Derivative to Descend



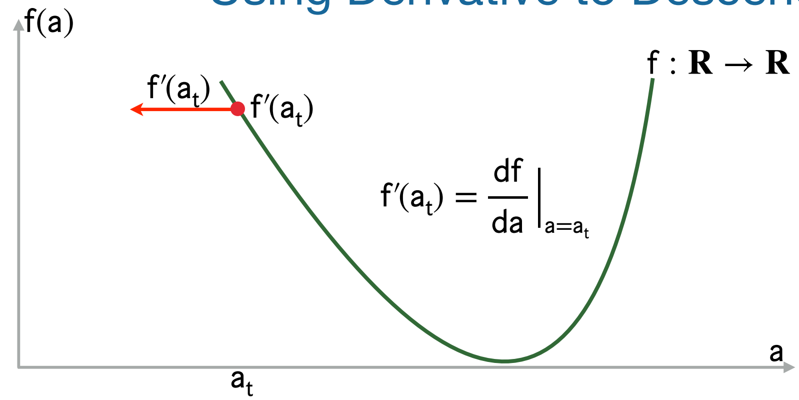
Using Derivative to Descend



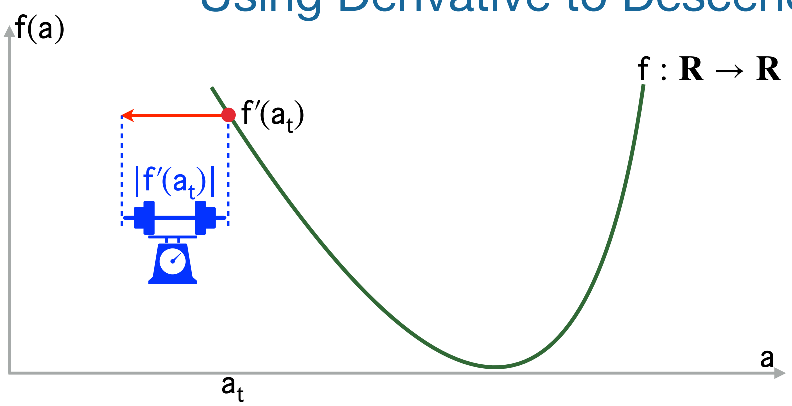
Using Derivative to Descend



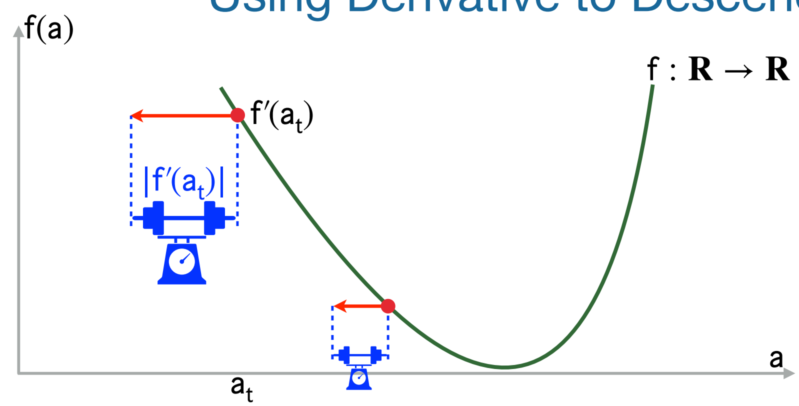
Using Derivative to Descend



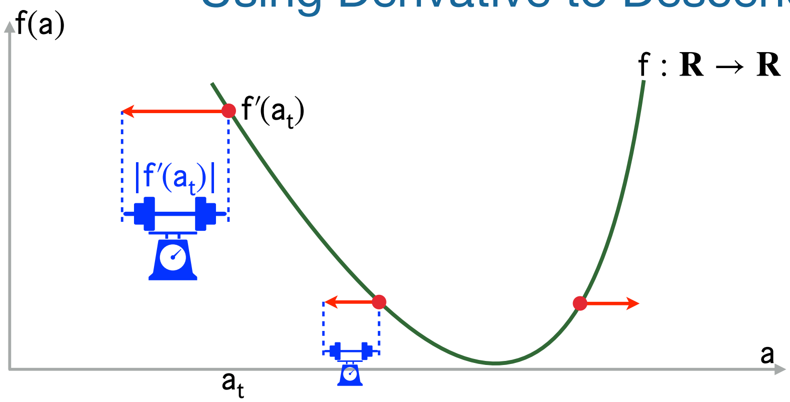
Using Derivative to Descend



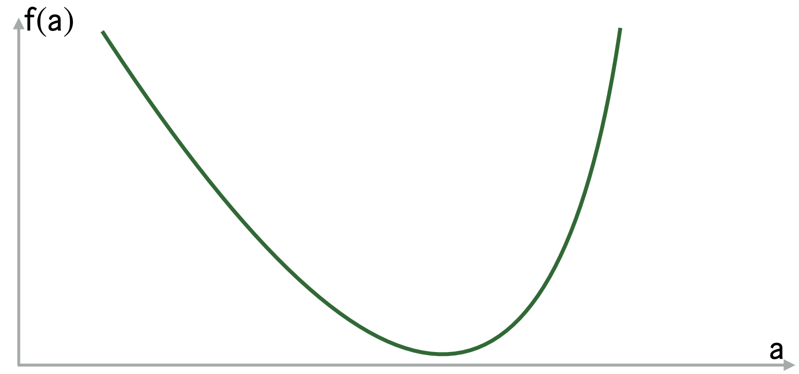
Using Derivative to Descend



Using Derivative to Descend

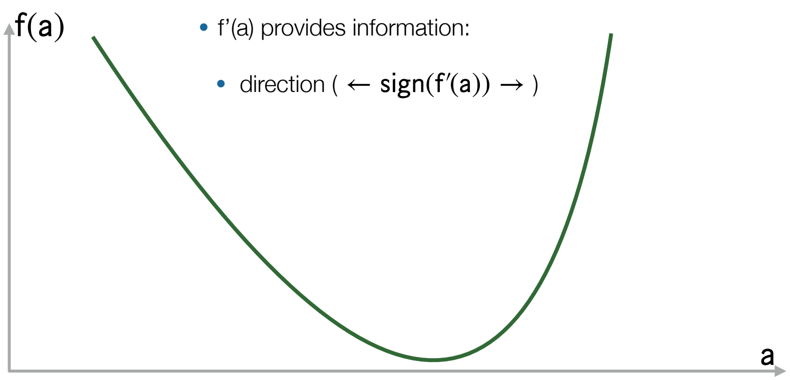


Iterative Derivative Procedure

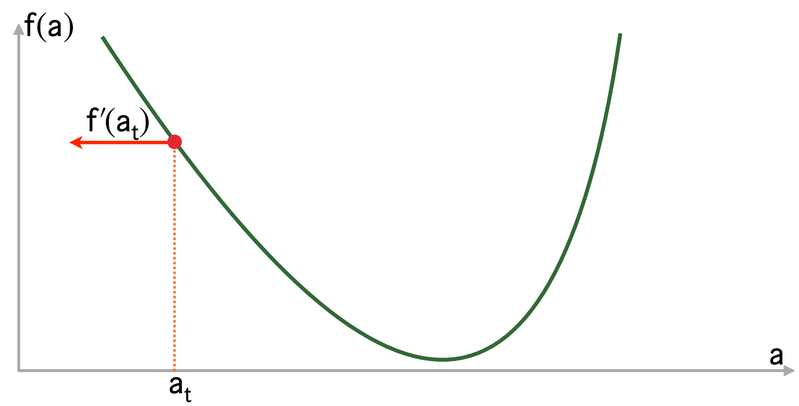


Iterative Derivative Procedure

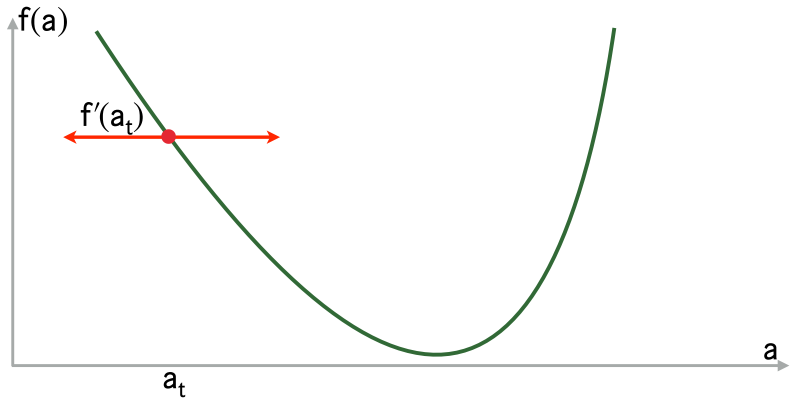
- f′(a) provides information:
  - direction (← sign(f′(a)) →)



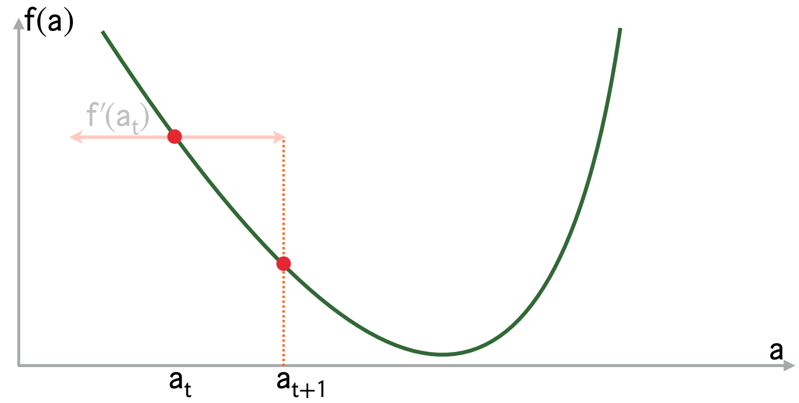
Iterative Derivative Procedure



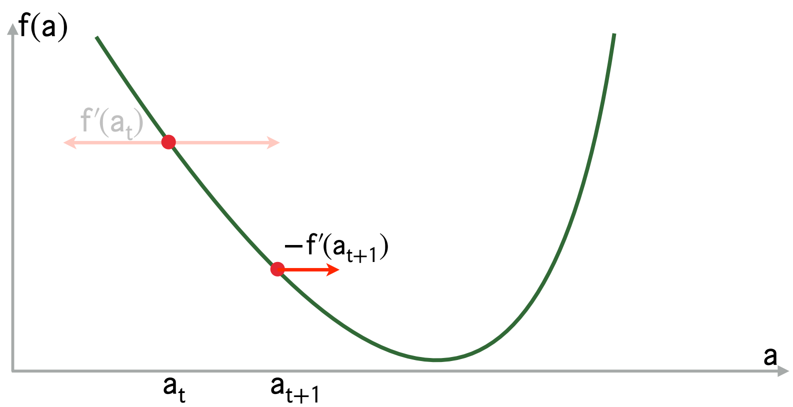
Iterative Derivative Procedure



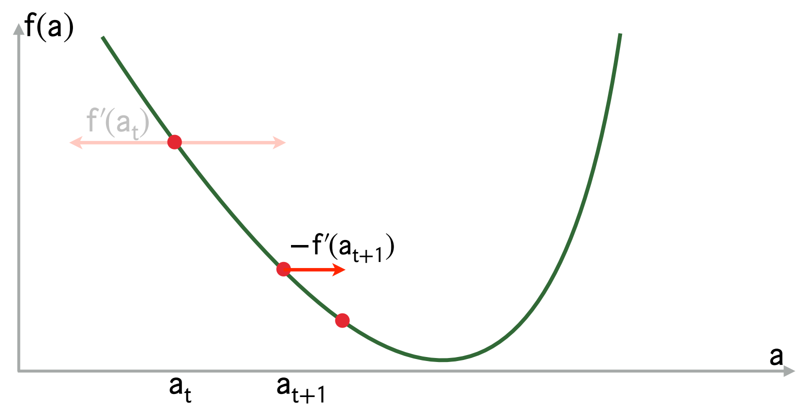
Iterative Derivative Procedure



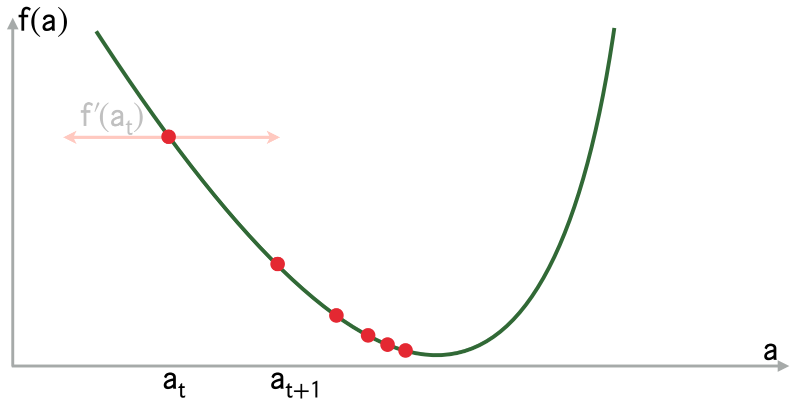
Iterative Derivative Procedure



Iterative Derivative Procedure



Iterative Derivative Procedure



Iterative Derivative Procedure

- Input: function f
- Goal: find $\hat{a}$ such that $|f(\hat{a}) - f(a)| \leq \epsilon$
- Choose initial value a_1
- Loop:
  - $a_{t+1} \leftarrow a_t - \eta f'(a_t)$
  - Until …

Iterative Derivative Procedure

- Input: function f
- Goal: find $\hat{a}$ such that $|f(\hat{a}) - f(a)| \leq \epsilon$
- Choose initial value a_1
- Loop:
  - $a_{t+1} \leftarrow a_t - \eta f'(a_t)$
  - Until …

⭐learning rate
⭐step size

Learning Rate

- Crucial in many learning problems
- Fixed learning-rate can be used in certain circumstances
- Self-tuning procedure of learning-rate exist, notably AdaGrad
- In many applications:
  - Linear decrease $\eta_t = \dfrac{\eta_0}{b + st}$ where $\eta_0 \in [0.1, 1]$,
  - Sub-linear decreas $\eta_t \sim \dfrac{\eta_0}{\sqrt{t}}$

Example

- Find minimum of

$$f(x) = \log\left(1 + e^{x-a-\delta}\right) + \log\left(1 + e^{b-a-\delta}\right)$$

- Derivative is

$$f(x) = \frac{1}{1 + e^{a+\delta-x}} - \frac{1}{1 + e^{x+\delta-b}}$$

- a=1 b=2 δ=4.5
- Run with different initializations and learning rates $\left(\eta \in \{0.01, 4, 40\}\right)$

Implementation of f(x)

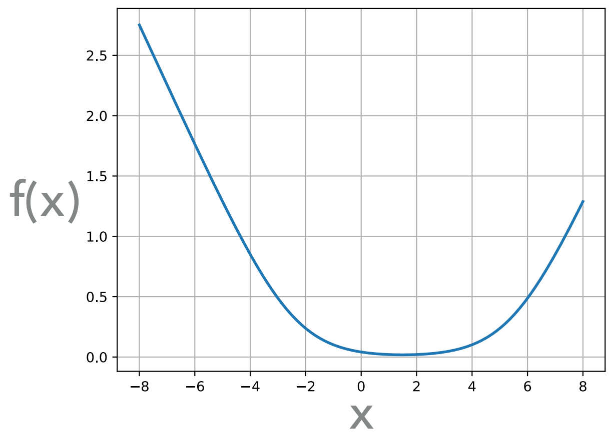
```
import numpy as np

def smooth_l1(a, b, delta):
    def smooth_l1_close(x):
        lloss = np.log(1 + np.exp(x - a - delta))
        rloss = np.log(1 + np.exp(b - x - delta))
        return 0.5 * (rloss + lloss)
    return smooth_l1_close
```

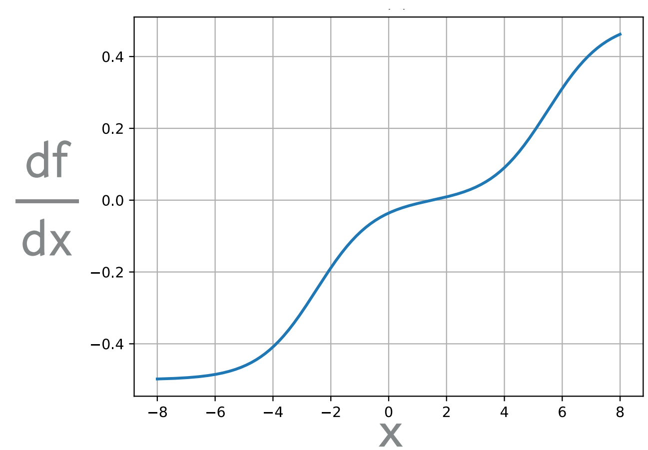

Implementation of f'(x)

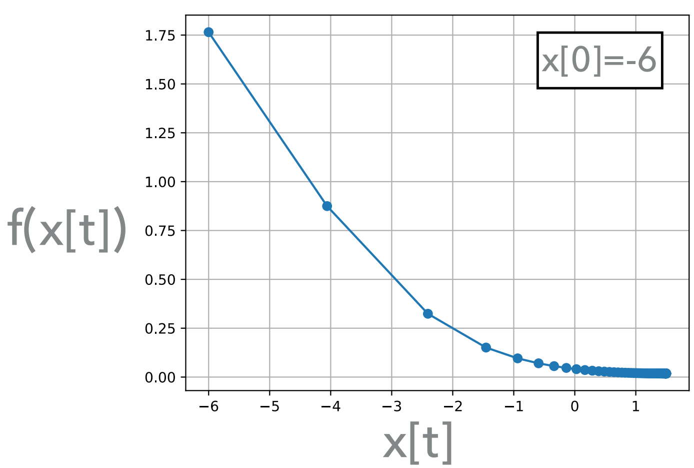
```
def smooth_l1_deriv(a, b, delta):
    def smooth_l1_deriv_close(x):
        rderiv = 1. / (1 + np.exp(a + delta - x))
        lderiv = 1. / (1 + np.exp(x + delta - b))
        return 0.5 * (rderiv - lderiv)
    return smooth_l1_deriv_close
```

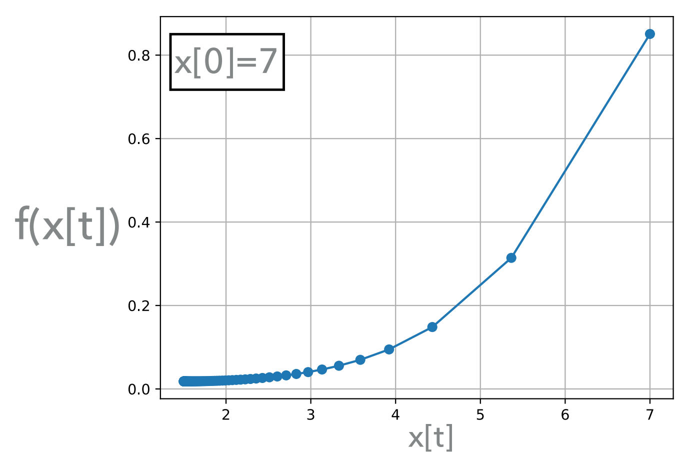

Derivative Descent

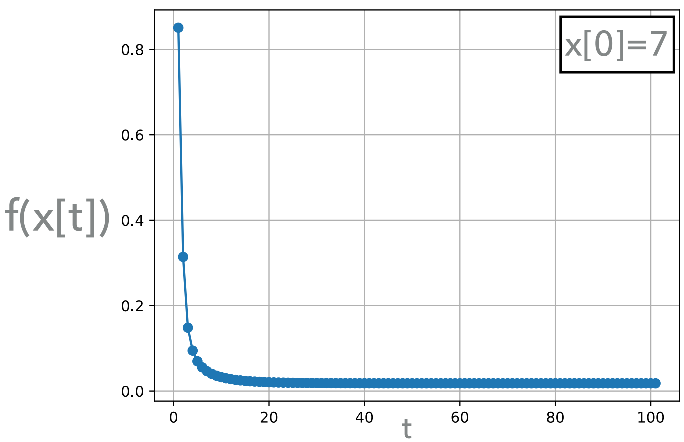
```
def derivative_descent(x0, deriv_func, eta):
    T = len(eta) + 1
    x = np.zeros(T)
    x[0] = x0
    for i in range(1, T):
        x[i] = x[i-1] - eta[i-1] deriv_func(x[i-1])
    return x
```

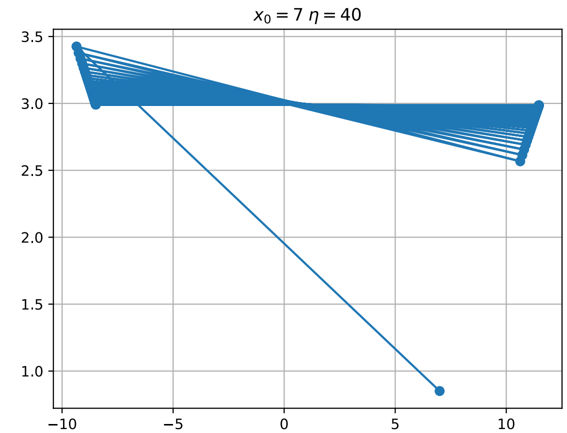

f(x)
x

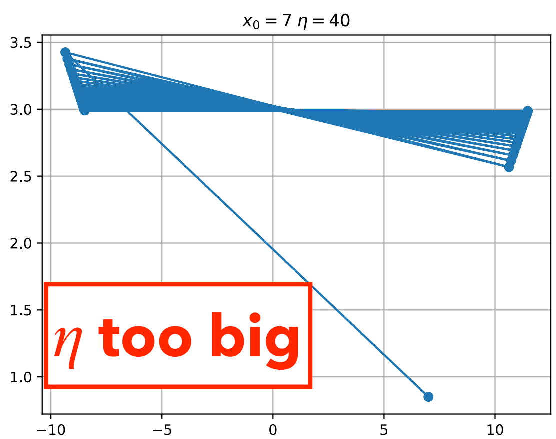

df
dx
x

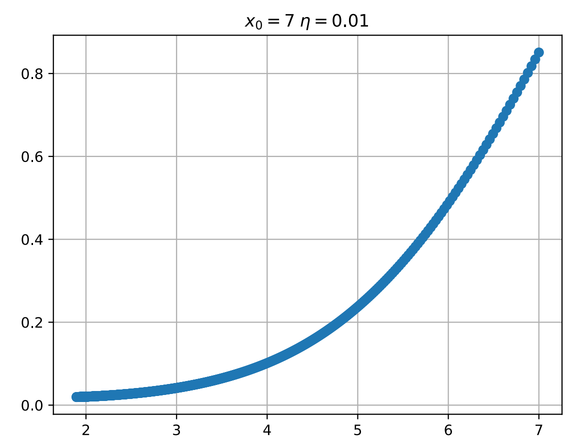

x[0]=-6
f(x[t])
x[t]

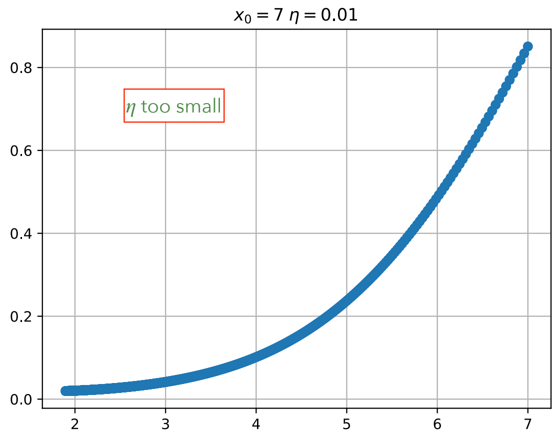

x[0]=7
f(x[t])
x[t]


x[0]=7
f(x[t])
t

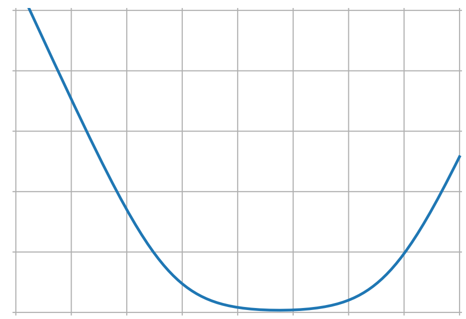

$x_0 = 7$ $\eta = 40$


$x_0 = 7$ $\eta = 40$
η too big

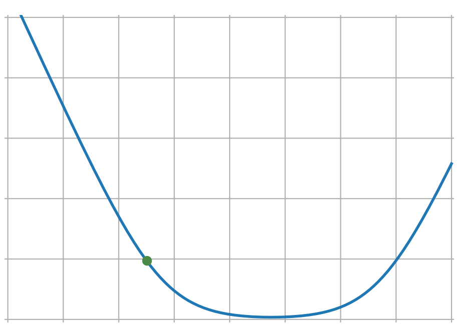

$x_0 = 7$ $\eta = 0.01$



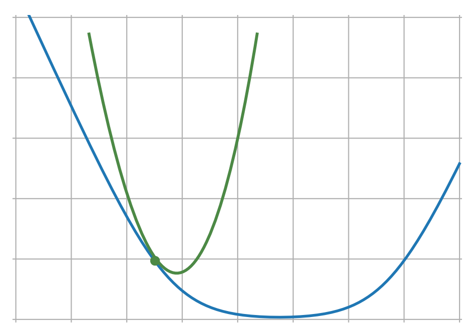
function f is called β-smooth: $f(x) \leq f(x_0) + f'(x_0)(x - x_0) + \frac{b}{2}(x - x_0)^2$



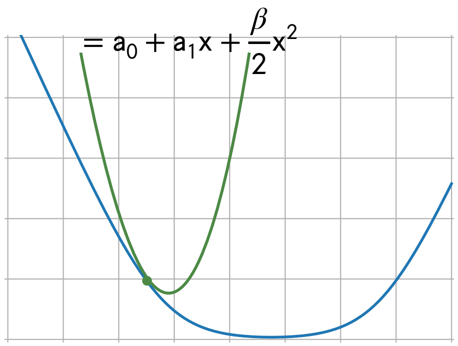
function f is called β-smooth: $f(x) \leq f(x_0) + f'(x_0)(x - x_0) + \frac{b}{2}(x - x_0)^2$



function f is called β-smooth: $f(x) \leq f(x_0) + f'(x_0)(x - x_0) + \frac{b}{2}(x - x_0)^2$



function f is called β-smooth: $f(x) \le f(x_0) + f'(x_0)(x - x_0) + \frac{b}{2}(x - x_0)^2$



function f is called β-smooth: $f(x) \le f(x_0) + f'(x_0)(x - x_0) + \frac{b}{2}(x - x_0)^2$

$$\text{f is } \beta\text{-smooth} \quad \Leftrightarrow \quad f''(x) \le \beta$$

DD for Smooth Functions

DD for Smooth Functions

- Set $\forall t: \; \eta_t = \frac{1}{\beta}$ which gives

$$x_{t+1} = x_t - \frac{1}{\beta} f'(x_t) = x_t - \frac{1}{\beta} g_t$$

DD for Smooth Functions

- Set $\forall t: \ \eta_t = \frac{1}{\beta}$ which gives

$$x_{t+1} = x_t - \frac{1}{\beta} f'(x_t) = x_t - \frac{1}{\beta} g_t$$

- From Smoothness:

$$f(x_{t+1}) \le f(x_t) + g_t (x_{t+1} - x_t) + \frac{\beta}{2}(x_{t+1} - x_t)^2$$

$$f(x_{t+1}) \le f(x_t) \ - \ \frac{1}{\beta} g_t^2 \ + \ \frac{\beta}{2} \frac{g_t^2}{\beta^2} = f(x_t) \ - \ \frac{1}{2\beta} g_t^2$$

DD for Smooth Functions

With $\eta_t = \frac{1}{\beta}$ and β-smoothness $f(x_{t+1}) \le f(x_t) \ - \ \frac{1}{2\beta} g_t^2$

DD for Smooth Functions

With $\eta_t = \frac{1}{\beta}$ and β-smoothness $f(x_{t+1}) \le f(x_t) \ - \ \frac{1}{2\beta} g_t^2$

However g_t gets smaller as we approach the minimum

DD for Smooth Functions

With $\eta_t = \frac{1}{\beta}$ and β-smoothness $f(x_{t+1}) \le f(x_t) \ - \ \frac{1}{2\beta} g_t^2$

However g_t gets smaller as we approach the minimum

It turns out that from convexity $g_t \ge \frac{\Delta_t}{|x_0 - x^\star|}$ where $\Delta_t = f(x_t) - f(x^\star)$

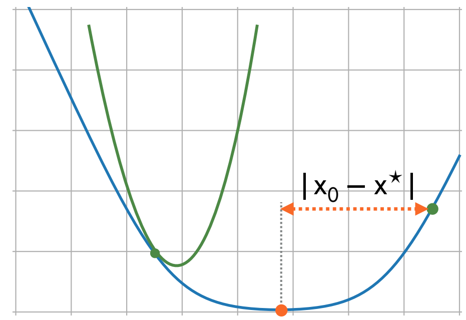
DD for Smooth Functions

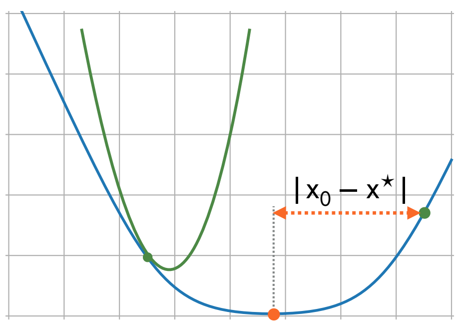
With $\eta_t = \dfrac{1}{\beta}$ and β-smoothness $f(x_{t+1}) \leq f(x_t) - \dfrac{1}{2\beta} g_t^2$

However $\mathbf{g_t}$ gets smaller as we approach the minimum

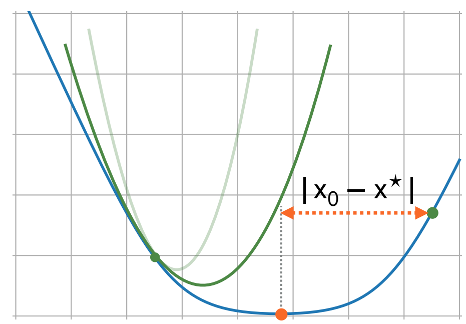
It turns out that from convexity $\mathbf{g_t} \geq \dfrac{\Delta_t}{|x_0 - x^\star|}$ where $\Delta_t = f(x_t) - f(x^\star)$

This gives $f(x_{t+1}) - f(x^\star) \leq \dfrac{2\beta |x_0 - x^\star|}{t}$



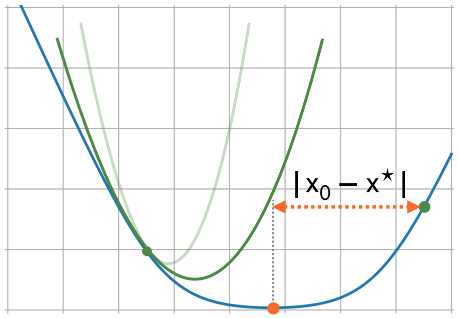


The closer to the optimum we start the faster we converge



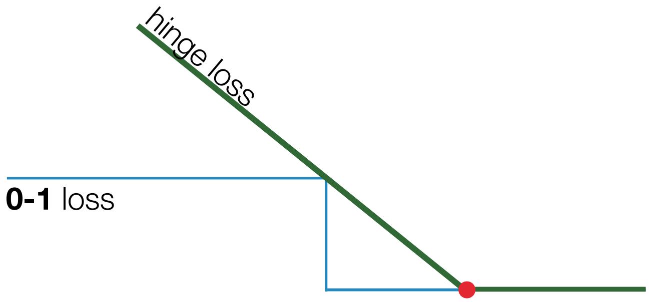
The closer to the optimum we start the faster we converge

The smaller β is the faster we convergence

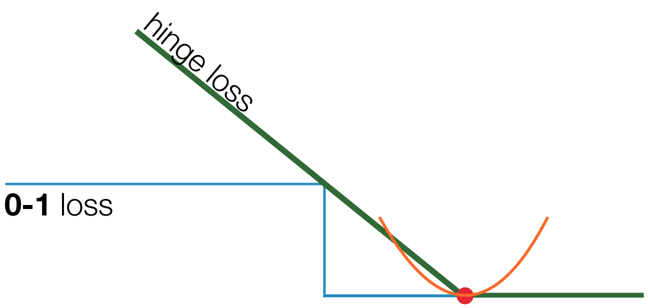


The closer to the optimum we start the faster we converge

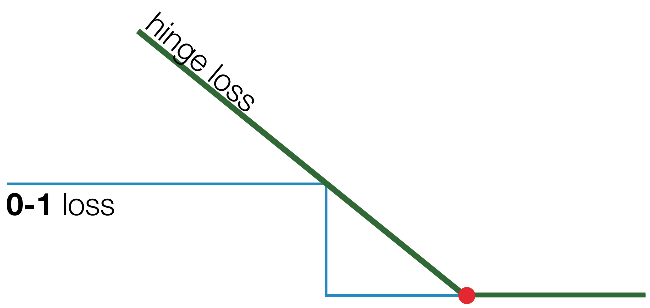
Non-smooth Functions



Non-smooth Functions

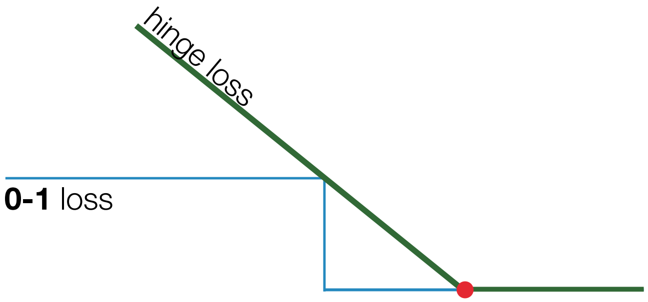


Non-smooth Functions



Non-smooth Functions

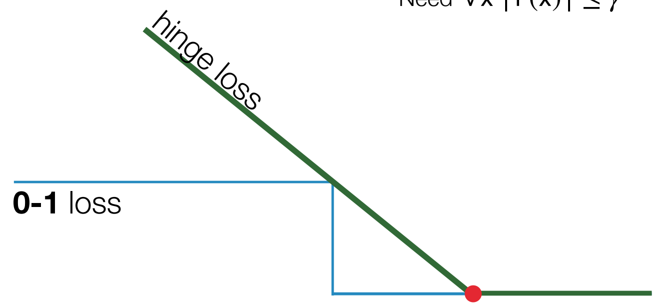
When loss function is not smooth:



Non-smooth Functions

When loss function is not smooth:

Need $\forall x \; |f'(x)| \le \gamma$

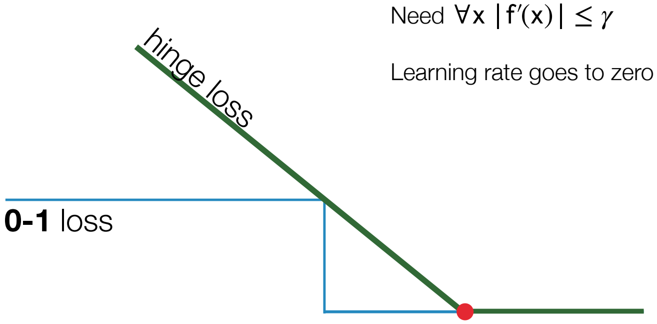


Non-smooth Functions

When loss function is not smooth:

Need $\forall x \; |f'(x)| \le \gamma$

Learning rate goes to zero $\eta_0 \, t^{-1/2}$



Non-smooth Functions

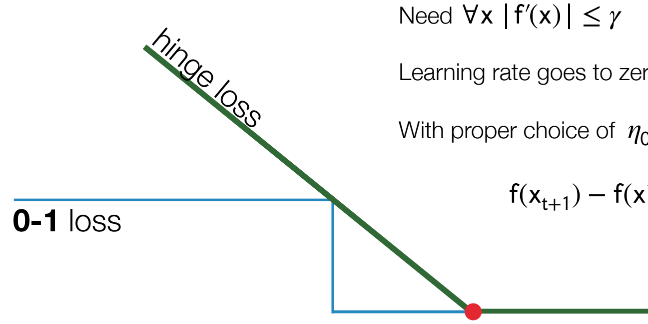
When loss function is not smooth:

Need $\forall x \; |f'(x)| \le \gamma$

Learning rate goes to zero $\eta_0 \, t^{-1/2}$

With proper choice of η_0

$$f(x_{t+1}) - f(x^\star) \le \frac{\gamma \, |x_0 - x^\star|}{\sqrt{t}}$$



Non-smooth Functions

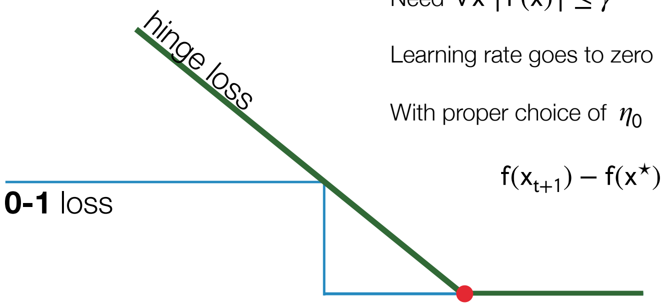
When loss function is not smooth:

Need $\forall x \; |f'(x)| \le \gamma$

Learning rate goes to zero $\eta_0 t^{-1/2}$

With proper choice of η_0

$$f(x_{t+1}) - f(x^\star) \le \frac{\gamma |x_0 - x^\star|}{\sqrt{t}}$$



Non-smooth Functions

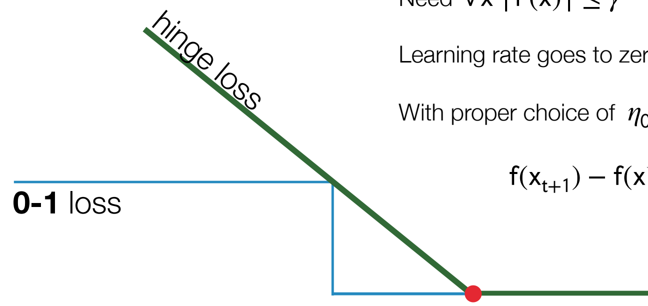
When loss function is not smooth:

Need $\forall x \; |f'(x)| \le \gamma$

Learning rate goes to zero $\eta_0 t^{-1/2}$

With proper choice of η_0

$$f(x_{t+1}) - f(x^\star) \le \frac{\gamma |x_0 - x^\star|}{\sqrt{t}}$$



© 2020 Yoram Singer

27

Next…

From one variable to a vector of variables

Gradients and their properties

Convexity of multivariate functions

Smoothness of multivariate functions

Gradient Descent

Stochastic Gradient Descent (SGD)

SGD for generalized linear models

SGD for non-linear models