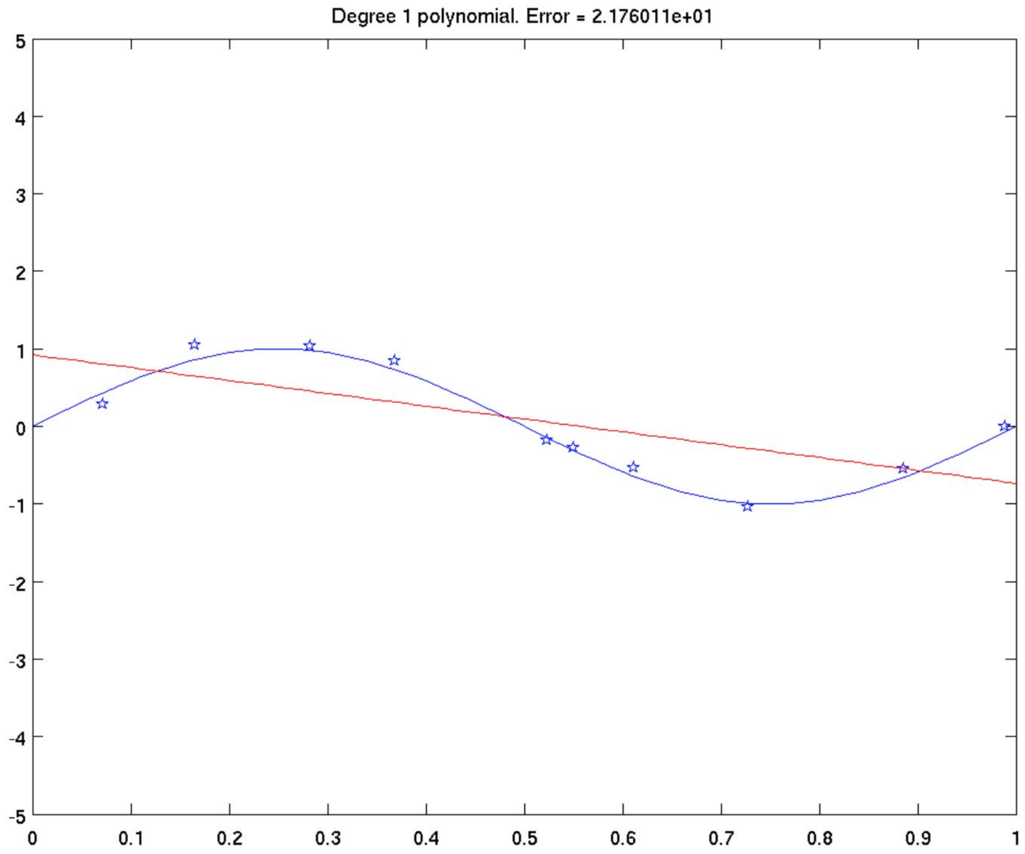
CMPSC 448: Machine Learning

Lecture 4. Basic Convex Optimization

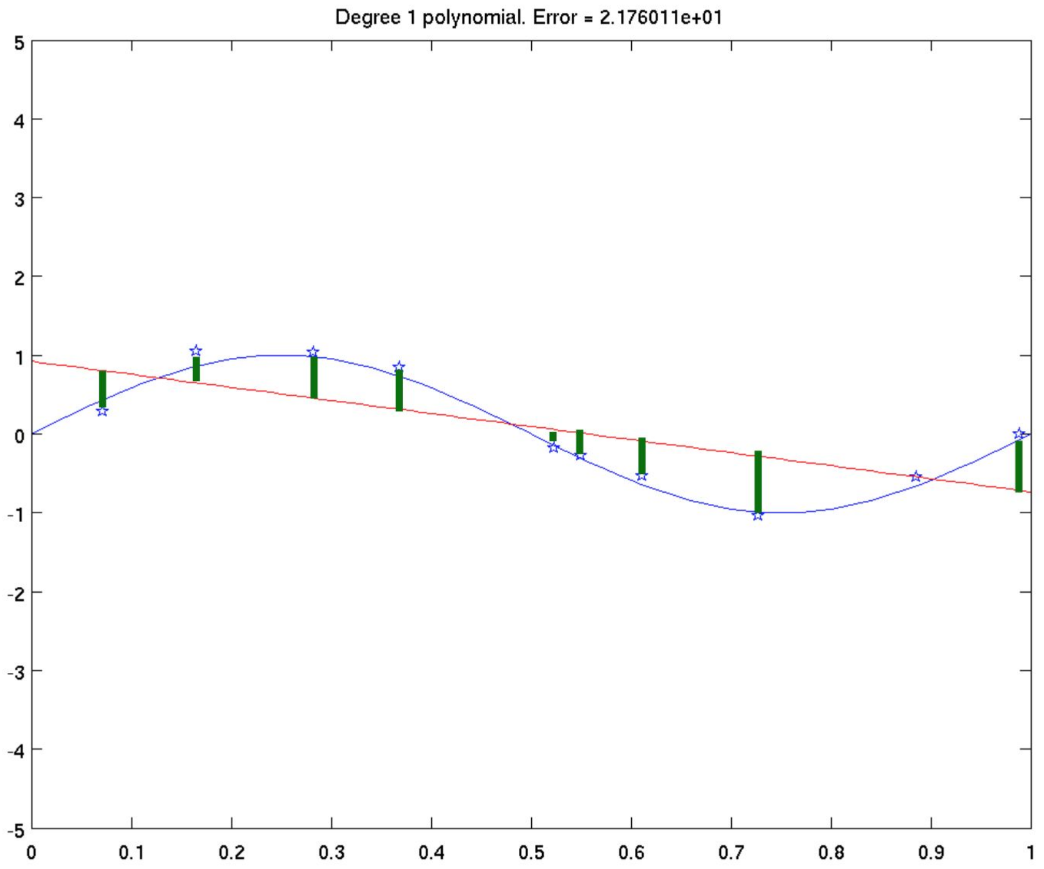
Rui Zhang
Fall 2021

PennState

Why optimization?



Why optimization?



Why optimization?

For a given training data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

$$\left(\begin{array}{c}\text{difference between } \textbf{true} \\ \textbf{value } \; y_i \; \text{ and } \textbf{what} \\ \textbf{model predicts for } x_i\end{array}\right) + \begin{array}{c}\textbf{some regularization} \\ \textbf{of model parameters}\end{array}$$

Why optimization?

For a given training data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

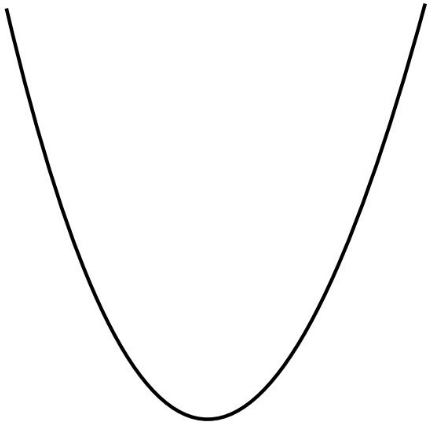
minimize

all possible values of model parameters

$$\sum_{i=1}^{n} \left(\text{difference between } \textbf{true value } y_i \text{ and } \textbf{what model predicts for } x_i \right) + \textbf{some regularization of model parameters}$$

One minute calculus

Find the minimum x_* of $f(x)$?

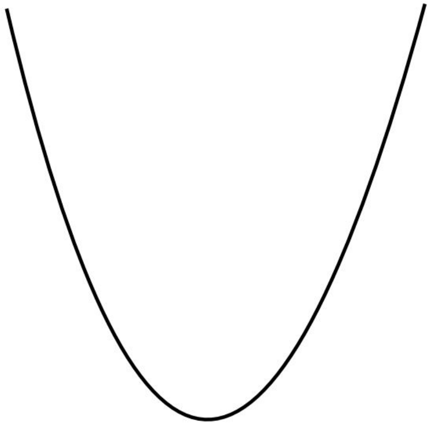


$$f(x) = (x-2)^2$$

One minute calculus

Find the minimum x_* of $f(x)$?

Easy: set the derivative to zero!



$$f(x) = (x-2)^2$$

$$f'(x) = 2(x-2) = 0$$

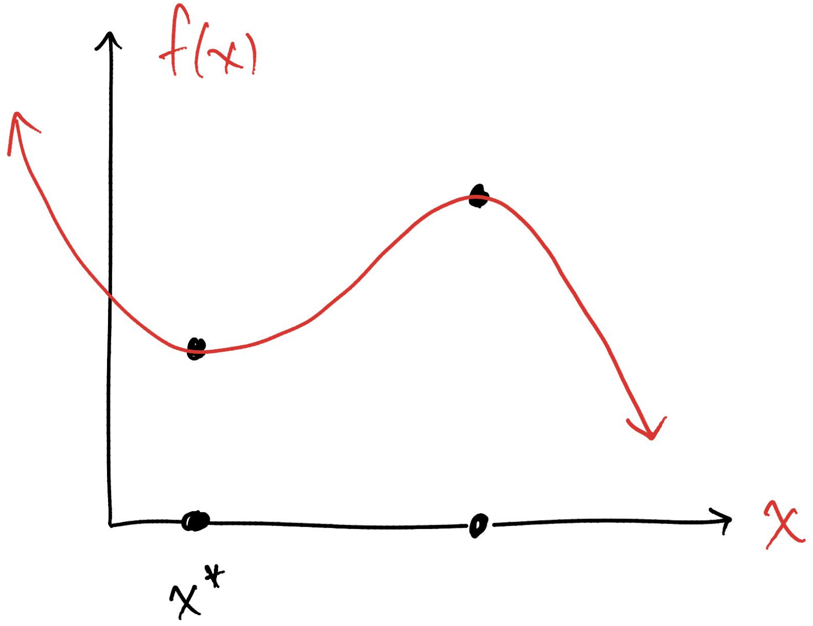
$$\Rightarrow \quad x_* = 2$$

Property 1

Theorem. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, if $f(\boldsymbol{x})$ is differentiable and and $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

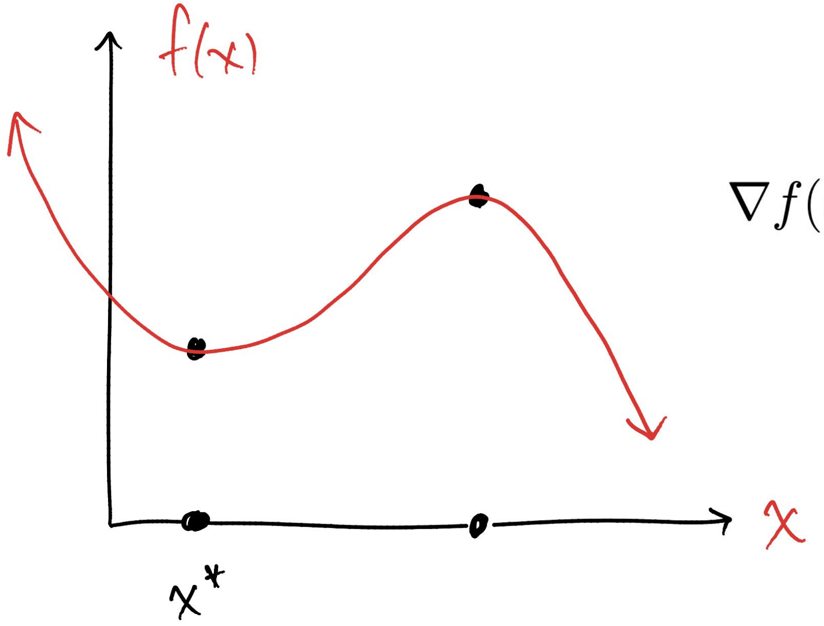
Property 1

Theorem. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, if $f(\boldsymbol{x})$ is differentiable and and $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.



Property 1

Theorem. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, if $f(\boldsymbol{x})$ is differentiable and and $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.



$\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is necessary but not sufficient

One minute calculus

How about this function?

$$\min_{x \in \mathbb{R}} \quad x^4 - 3x^3 + x^2 + \frac{3}{2}x$$

$$f'(x) = \frac{df}{dx} = 4x^3 - 9x^2 + 2x + \frac{3}{2}$$

One minute calculus

How about this function?

$$\min_{x \in \mathbb{R}} \quad x^4 - 3x^3 + x^2 + \frac{3}{2}x$$

$$f'(x) = \frac{df}{dx} = 4x^3 - 9x^2 + 2x + \frac{3}{2}$$

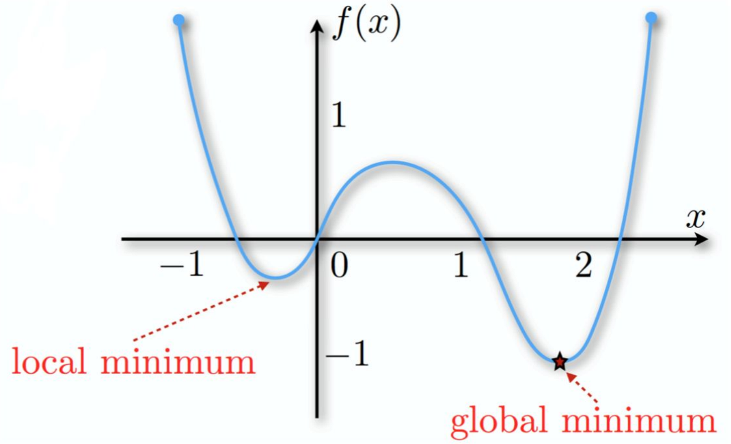
(1) There might not be a closed form solution for $f'(x) = 0$!

One minute calculus

How about this function?

$$\min_{x\in\mathbb{R}} \quad x^4 - 3x^3 + x^2 + \frac{3}{2}x$$

$$f'(x) = \frac{df}{dx} = 4x^3 - 9x^2 + 2x + \frac{3}{2}$$



(1) There might not be a closed form solution for $f'(x) = 0$!

(2) Having derivative equals zero is NOT sufficient for optimality!

Property 2

Theorem. If $f(\boldsymbol{x})$ is twice continuously differentiable and $\boldsymbol{x}^*$ is a local minimum, then $\nabla^2 f(\boldsymbol{x}^*)$ is positive semidefinite (i.e., $z^\intercal \nabla^2 f(\boldsymbol{x}^*) z \geq 0, \quad \forall z \in \mathbb{R}^d$).

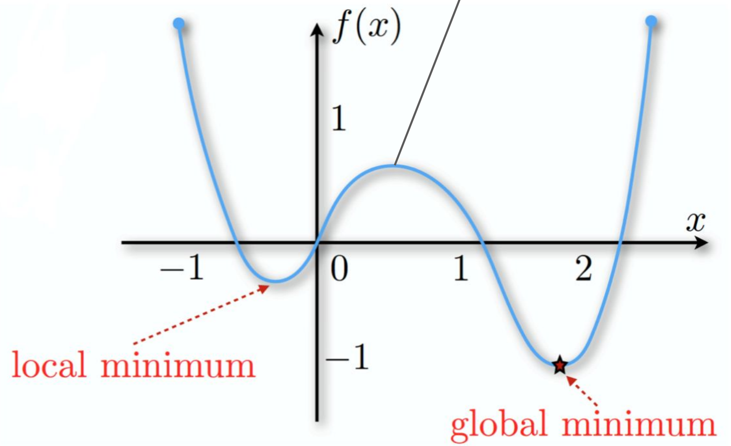
Property 2

Theorem. If $f(\boldsymbol{x})$ is twice continuously differentiable and $\boldsymbol{x}^*$ is a local minimum, then $\nabla^2 f(\boldsymbol{x}^*)$ is positive semidefinite (i.e., $z^\intercal \nabla^2 f(\boldsymbol{x}^*) z \geq 0, \quad \forall z \in \mathbb{R}^d$).

This can't be a local minimum because second order derivative <0

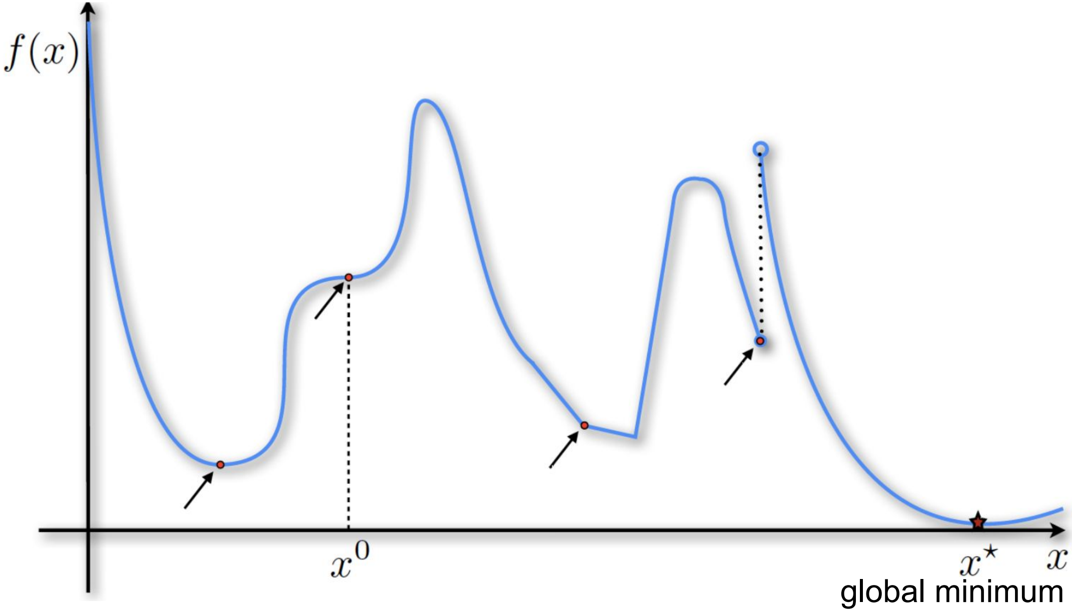
$$\min_{x \in \mathbb{R}} \; x^4 - 3x^3 + x^2 + \frac{3}{2}x$$

$$f'(x) = \frac{df}{dx} = 4x^3 - 9x^2 + 2x + \frac{3}{2}$$



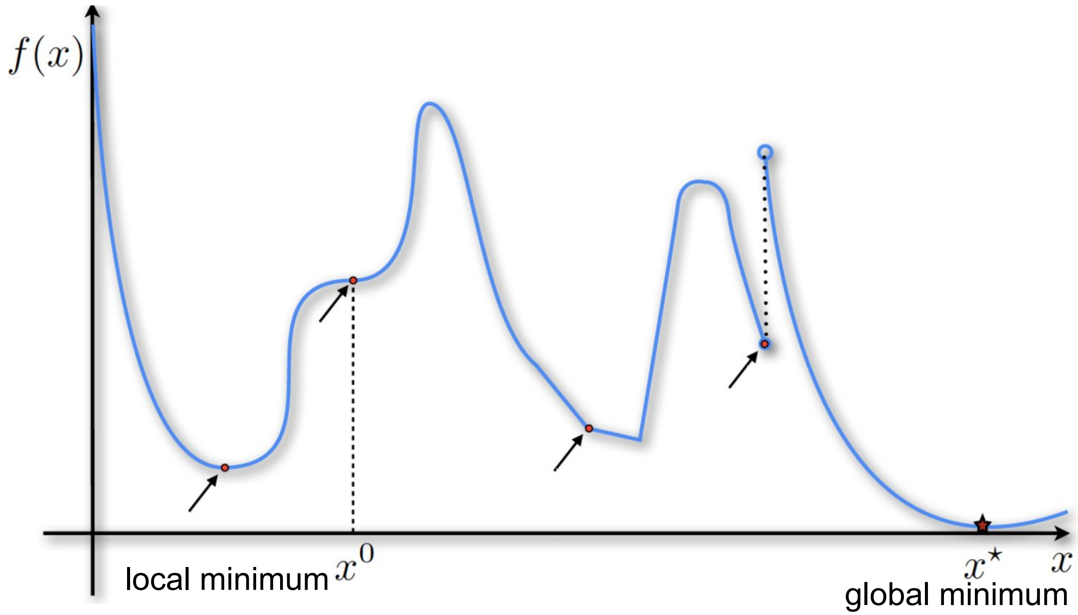
Iterative optimization

Fog of war



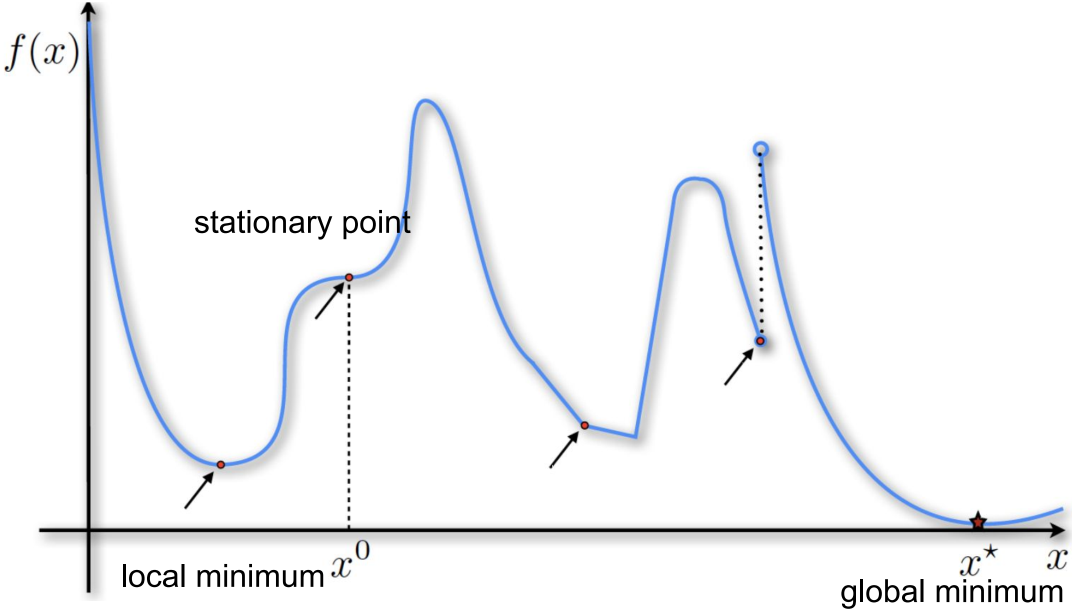
Iterative optimization

Fog of war



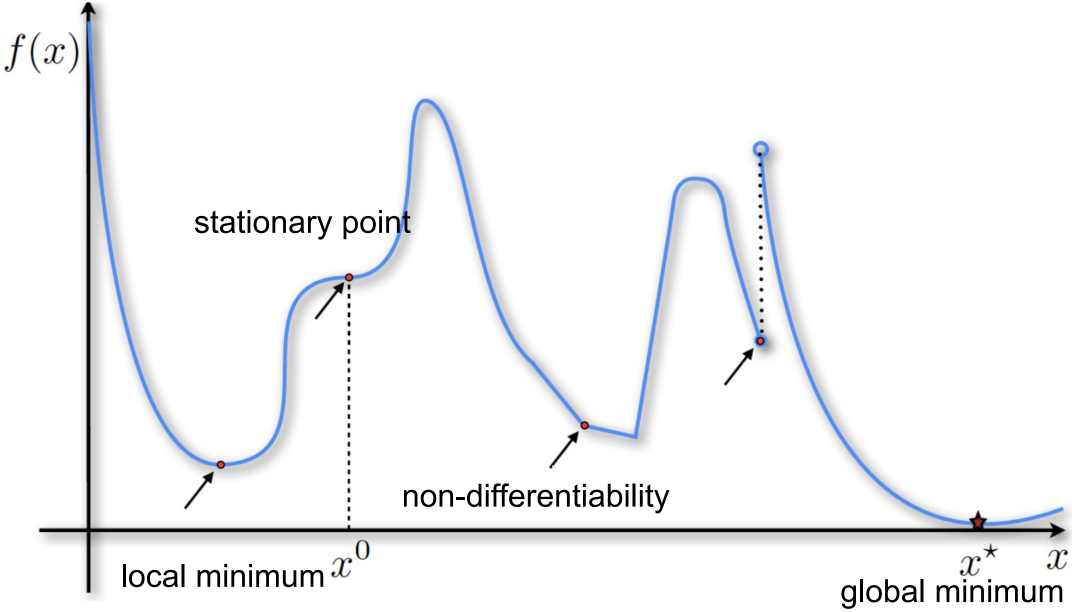
Iterative optimization

Fog of war



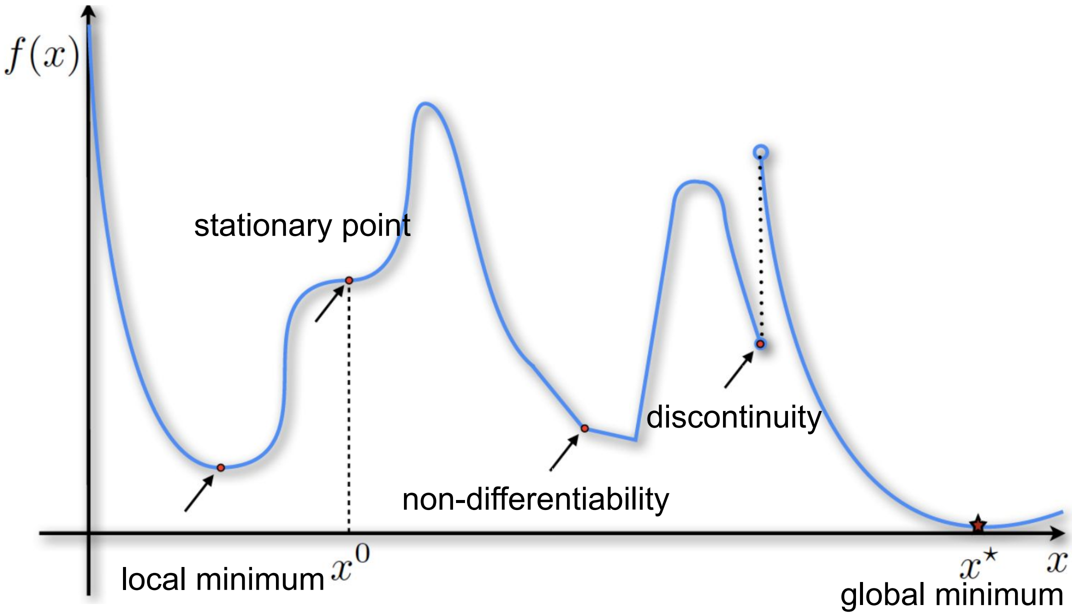
Iterative optimization

Fog of war



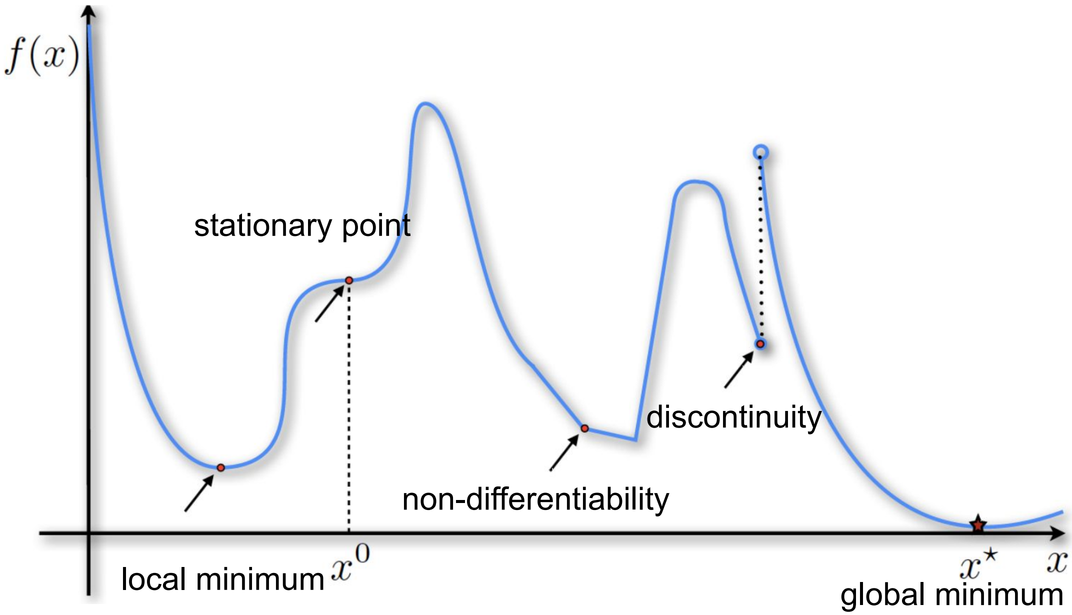
Iterative optimization

Fog of war



Iterative optimization

Fog of war



We need a key structure on the function local minima: Convexity.

Convex set

Definition

A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and any $\lambda \in [0, 1]$, we have:

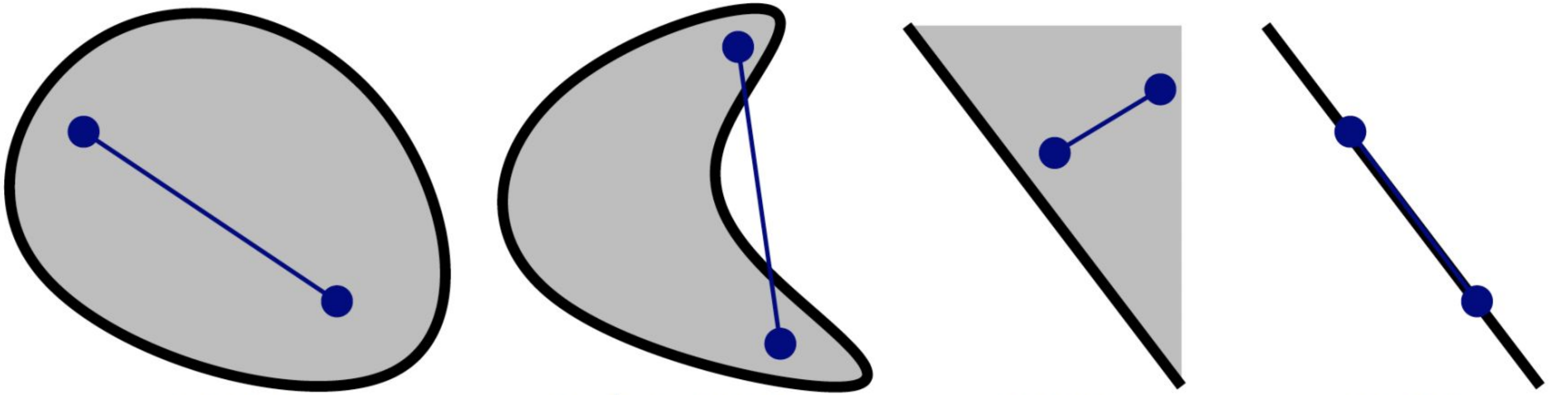
$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}$$

Convex set

Definition

A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and any $\lambda \in [0, 1]$, we have:

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}$$

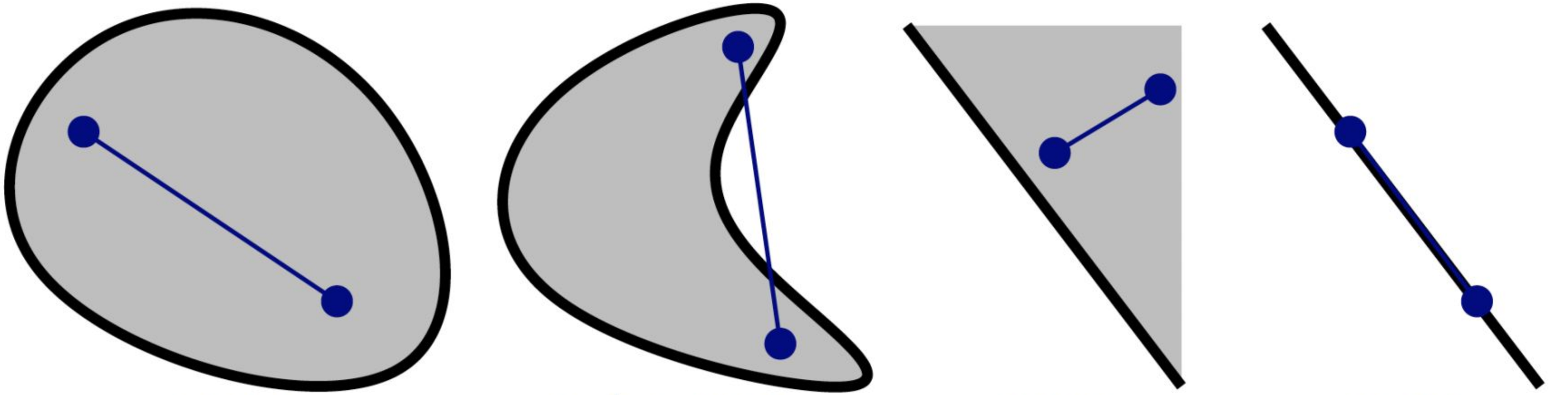


Convex set

Definition

A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and any $\lambda \in [0, 1]$, we have:

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}$$



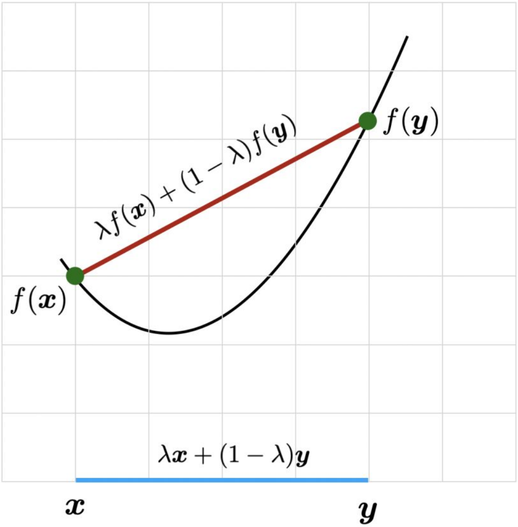
convex not convex convex convex

Convex function

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ if and only if:

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$

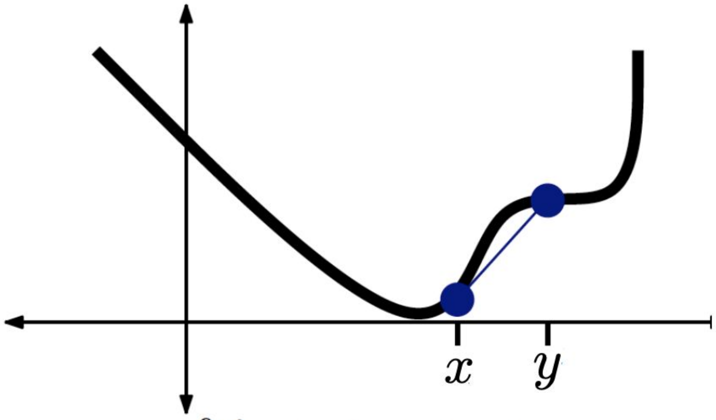


Convex function

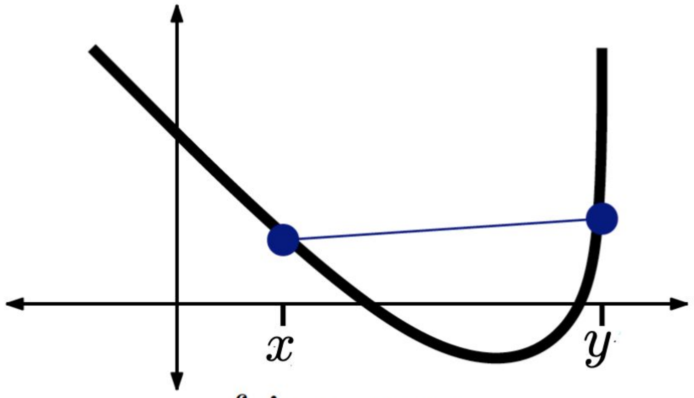
Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ if and only if:

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$$



f is not convex

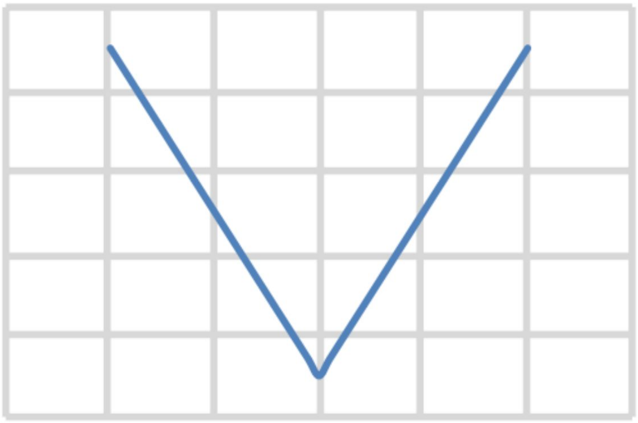


f is convex

Property 3

Theorem. If $f(\boldsymbol{x})$ is convex, then every local minimum is a global minimum.

Example: absolute



$f(x) = |x|$

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y| \\ &\leq |\lambda x| + |(1-\lambda)y| \\ &= \lambda|x| + (1-\lambda)|y| \\ &= \lambda f(x) + (1-\lambda)f(y) \end{aligned}$$

Example: norm

Is $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$ convex for $\boldsymbol{x} \in \mathbb{R}^d$?

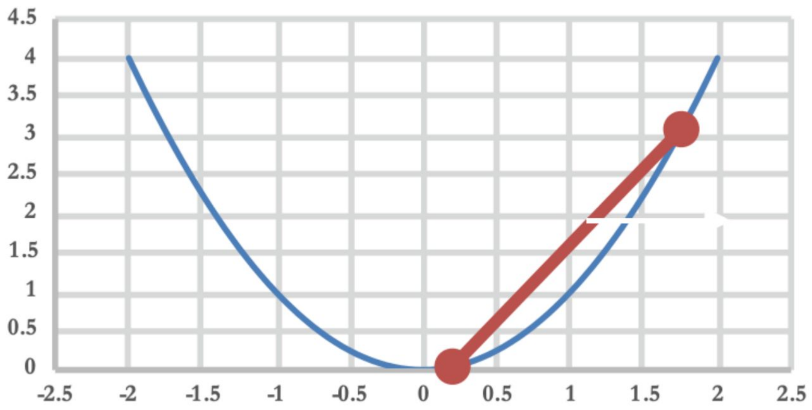
Example: norm

Is $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$ convex for $\boldsymbol{x} \in \mathbb{R}^d$?

$$
\begin{aligned}
f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &= \|\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\|_2 \\
&\leq \|\lambda \boldsymbol{x}\|_2 + \|(1-\lambda)\boldsymbol{y}\|_2 \quad \textbf{(triangle inequality)} \\
&= \lambda \|\boldsymbol{x}\|_2 + (1-\lambda)\|\boldsymbol{y}\|_2 \quad \textbf{(homogeneity)} \\
&= \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})
\end{aligned}
$$

Yes, the norm of a vector is a convex function.

Example: quadratic

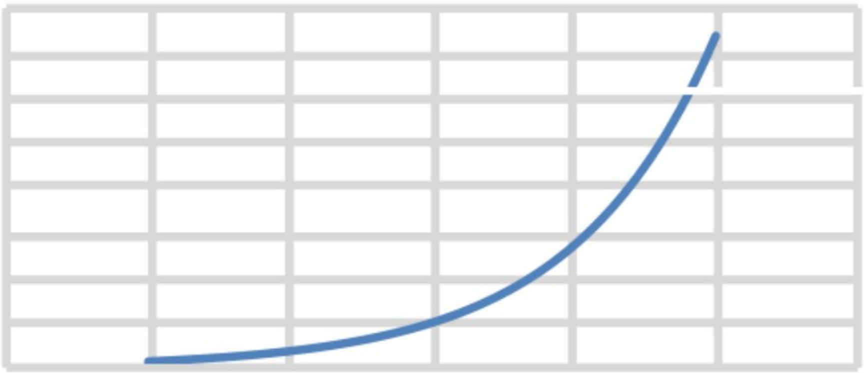


$f(x) = x^2$

$$
\begin{aligned}
\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda) y) &= \lambda x^2 + (1-\lambda) y^2 - (\lambda x + (1-\lambda) y)^2 \\
&= \lambda x^2 + (1-\lambda) y^2 - \lambda^2 x^2 - 2\lambda(1-\lambda) xy - (1-\lambda)^2 y^2 \\
&= \lambda(1-\lambda) x^2 + \lambda(1-\lambda) y^2 - 2\lambda(1-\lambda) xy \\
&= \lambda(1-\lambda)(x^2 + y^2 - 2xy) \\
&= \lambda(1-\lambda)(x-y)^2 \geq 0
\end{aligned}
$$

Example: exponential

$$f(x) = \exp(x) = e^x$$



- Show that above function is convex using basic definition of convexity?
- While it is obviously convex, You will find it's a bit hard to prove…
- but we can use second order derivative to show this very easy, which will be discussed later

Property 4: Alternate Definition of Convex Functions

Theorem

Let f be a differentiable function. Then, f is convex if and only if its domain is convex and the following inequalities hold:

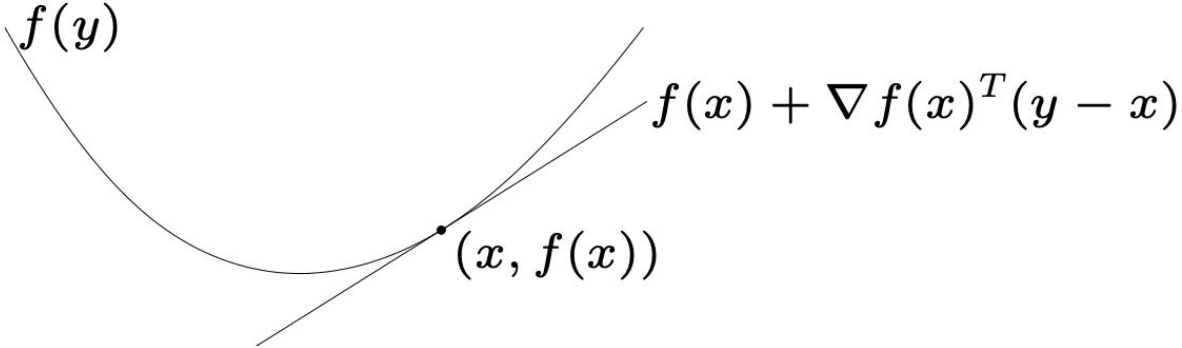
$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

Property 4: Alternate Definition of Convex Functions

Theorem

Let f be a differentiable function. Then, f is convex if and only if its domain is convex and the following inequalities hold:

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$



Property 4: Alternate Definition of Convex Functions

Is $f(\boldsymbol{x}) = e^{\boldsymbol{x}^\top \boldsymbol{a}}$ convex ?

$$
\begin{aligned}
f(\boldsymbol{y}) - (f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle) &= e^{\langle \boldsymbol{y}, \boldsymbol{a} \rangle} - \left(e^{\langle \boldsymbol{x}, \boldsymbol{a} \rangle} + e^{\langle \boldsymbol{x}, \boldsymbol{a} \rangle} \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{a} \rangle \right) \\
&= e^{\langle \boldsymbol{x}, \boldsymbol{a} \rangle} \left(e^{\langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{a} \rangle} - (1 + \langle \boldsymbol{a}, \boldsymbol{y} - \boldsymbol{x} \rangle) \right) \\
&\geq 0 \quad (\text{because} \quad 1 + z \leq e^z \quad \text{for all} \quad z \in \mathbb{R})
\end{aligned}
$$

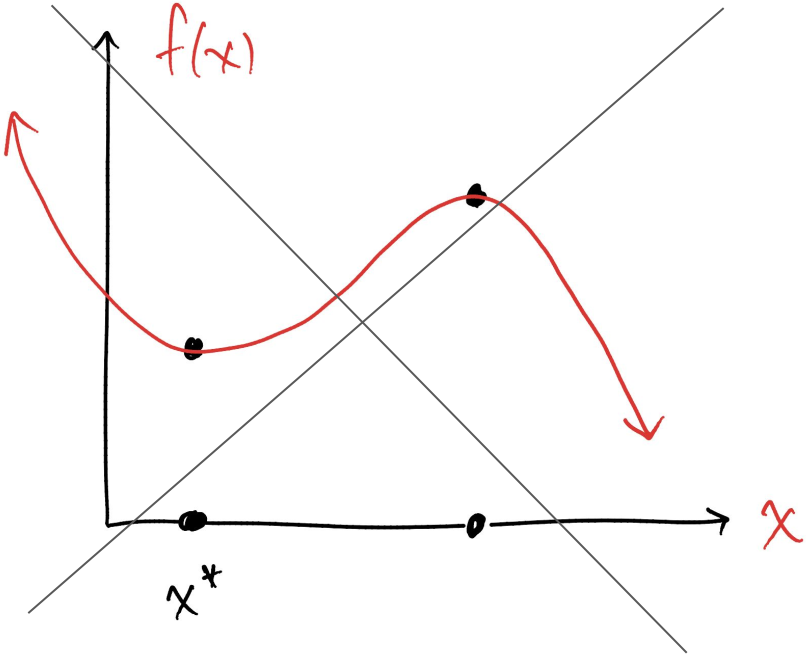
Yes, it is!

Property 5: Optimality condition for Convex function

Theorem. If $f(\boldsymbol{x})$ is convex and continuously differentiable, then $\boldsymbol{x}^*$ is a global minimum if and only if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Property 5: Optimality condition for Convex function

Theorem. If $f(\boldsymbol{x})$ is convex and continuously differentiable, then $\boldsymbol{x}^*$ is a global minimum if and only if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.



When convex, $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is necessary and sufficient

Property 5: Optimality condition for Convex function

Theorem. If $f(\boldsymbol{x})$ is convex and continuously differentiable, then $\boldsymbol{x}^*$ is a global minimum if and only if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Why?

From Property 4 we have:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}_*) + \langle \nabla f(\boldsymbol{x}_*), \boldsymbol{y} - \boldsymbol{x}_* \rangle$$

When $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}_*)$$

Property 6: twice continuously differentiable functions

- If the function $f : \mathbb{R} \to \mathbb{R}$ is twice-differentiable, then it is convex if and only if:

$$f''(x) \geq 0$$

Property 6: twice continuously differentiable functions

Is $f(x) = x^4$ convex ?

$$f''(x) = 12x^2 \geq 0$$

Yes, it is!

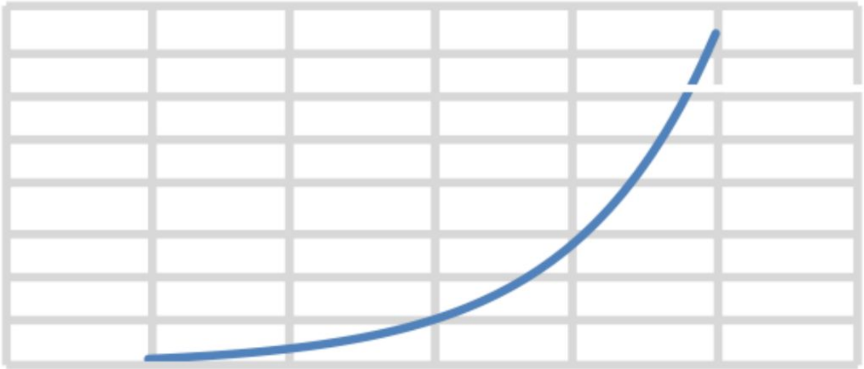
Property 6: twice continuously differentiable functions

Is $f(x) = x^4$ convex ?

$$f''(x) = 12x^2 \geq 0$$

Yes, it is!

$$f(x) = \exp(x) = e^x$$



Property 6: twice continuously differentiable functions

- If the function $f : \mathbb{R} \rightarrow \mathbb{R}$ is twice-differentiable, then it is convex if and only if:

$$f''(x) \geq 0$$

- If the function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is twice-differentiable, then it is convex if and only if:

$$\nabla^2 f(\boldsymbol{x}) \succeq 0$$

for all $\boldsymbol{x} \in \mathbb{R}^d$

- the Hessian matrix is positive semidefinite
- all the eigenvalues of its Hessian matrix are non-negative

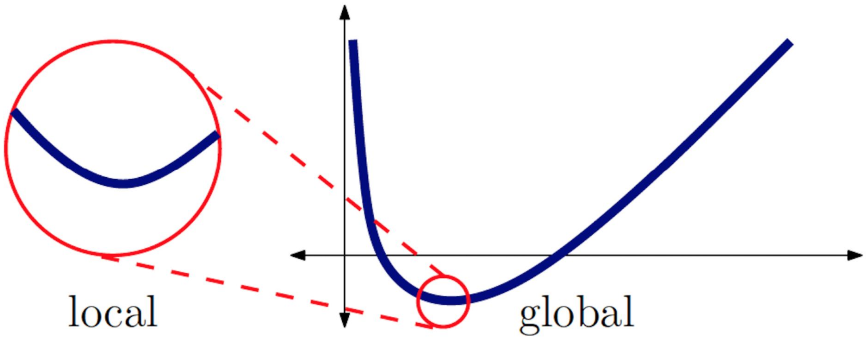
Convex Optimization

Problem:

Find the minimum of x_* of $f(x)$, when the function is convex!

From Property 3:

Every local minimum is a global minimum for convex functions!

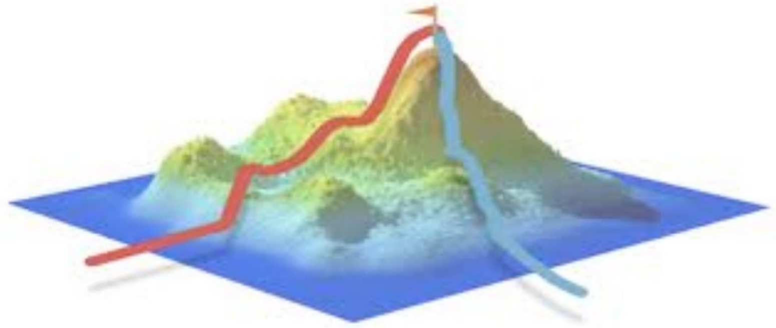


Convex functions are EASY to solve!

It suffices to find a local minimum, because we know it will be global

Descent direction

Let assume at iteration t the algorithms is at point $\boldsymbol{x}_t$ and got local information from oracle such as $f(\boldsymbol{x}_t)$ and $\nabla f(\boldsymbol{x}_t)$



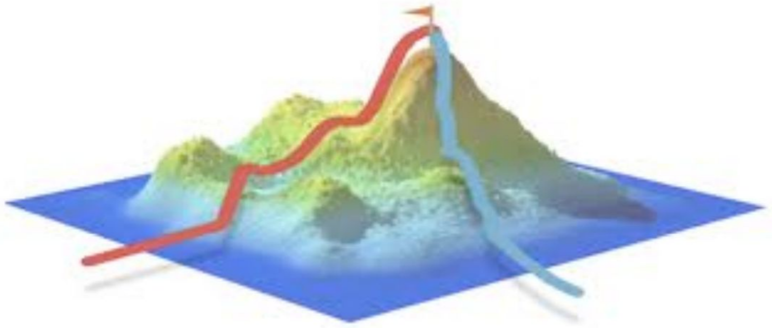
I would like to move to a new point $\boldsymbol{x}_{t+1}$

such that

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t)$$

Descent direction

Let assume at iteration t the algorithms is at point $\boldsymbol{x}_t$ and got local information from oracle such as $f(\boldsymbol{x}_t)$ and $\nabla f(\boldsymbol{x}_t)$



I would like to move to a new point $\boldsymbol{x}_{t+1}$

such that

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t)$$

Answer? negative gradient at current point $-\nabla f(\boldsymbol{x}_t)$

Gradient Descent (GD) algorithm

The simplest algorithm in the world (almost)

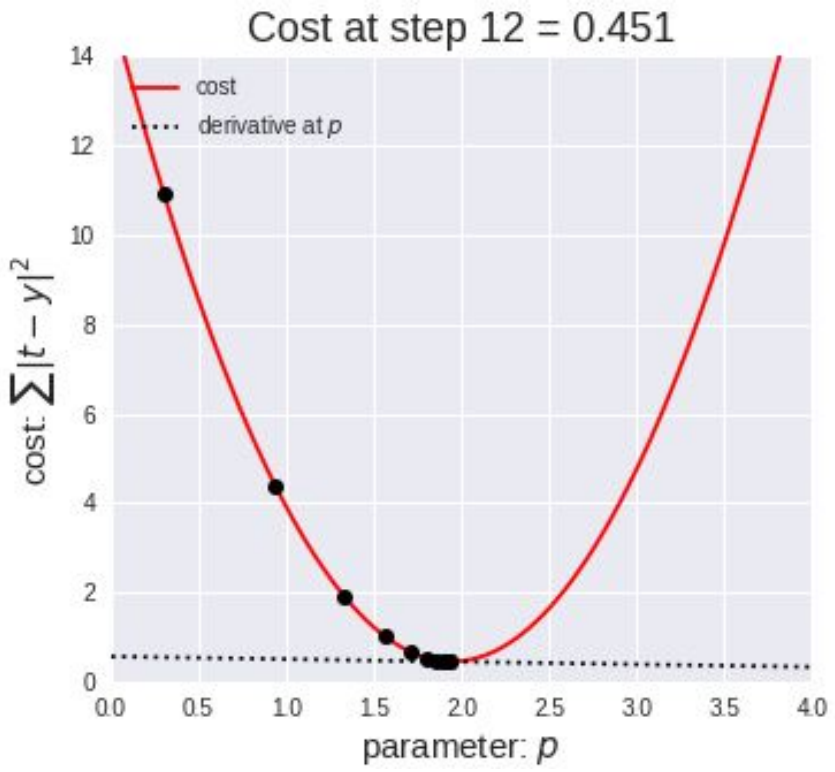
Initialize $\quad \boldsymbol{x}_0$

Iterate $\quad t = 1, 2, \ldots, T$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$

Step size

Example



Cost at step 12 = 0.451
cost
derivative at p
cost: $\sum|t-y|^2$
parameter: p
14
12
10
8
6
4
2
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Step size selection

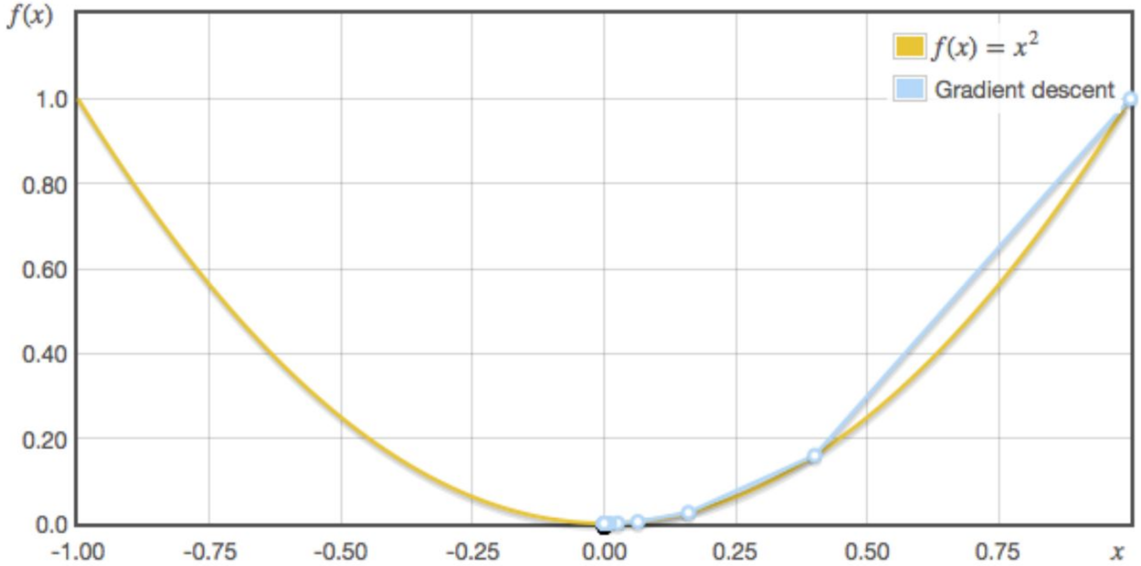
How do I choose the step size?

- Exact line search (usually expensive)
- Heuristics (practical)
- Fixed
- Adaptive based on iteration # [smaller steps at end]

Example Step size selection

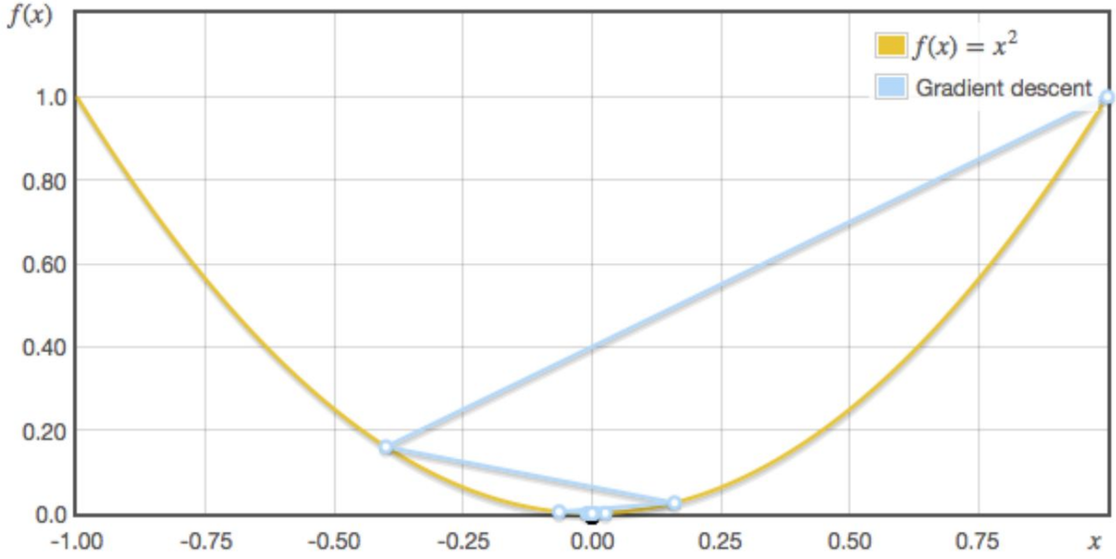
$$f(x) = x^2$$

$$\eta = 0.3$$



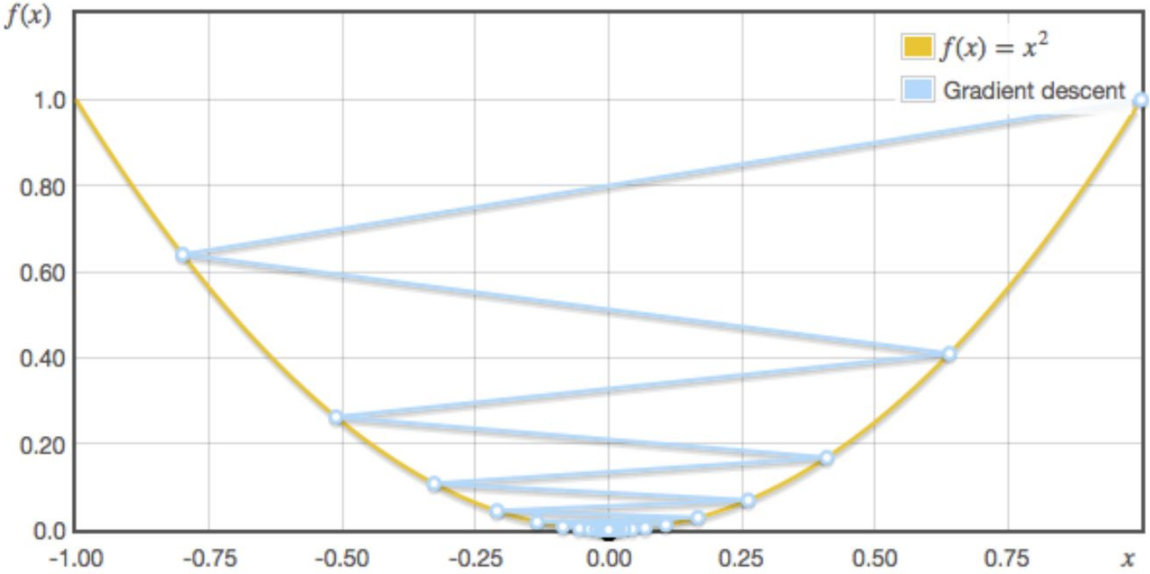
Example Step size selection

$$f(x) = x^2$$
$$\eta = 0.7$$



Example Step size selection

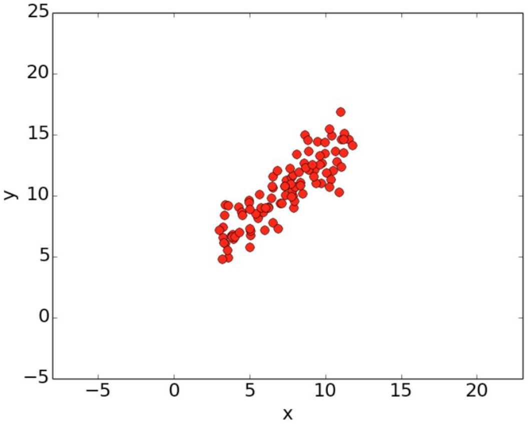
$$f(x) = x^2$$
$$\eta = 0.9$$



Polynomial degree 1

Given: a set of points on the plane

Goal: find the best line that approximates the points

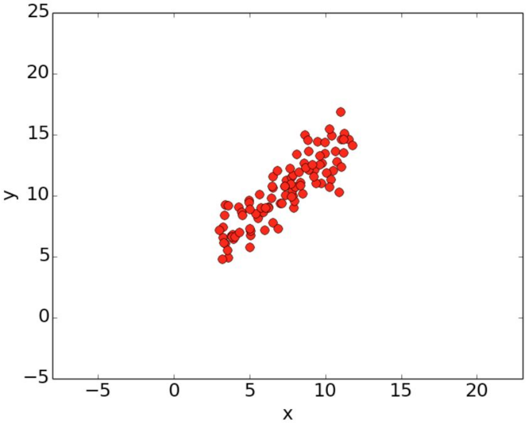


$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots$$
$$(x_n, y_n)$$

Polynomial degree 1

Given: a set of points on the plane

Goal: find the best line that approximates the points

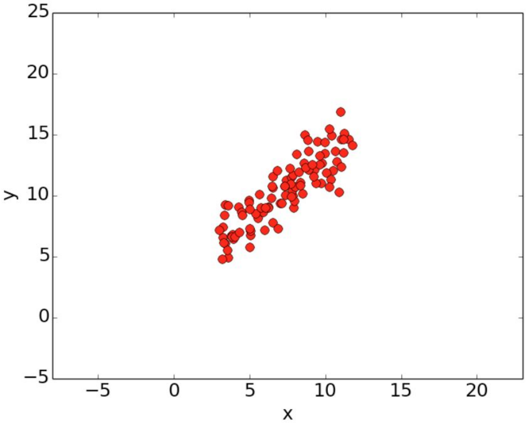


$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots$$
$$(x_n, y_n)$$

Error of a line:

$$f(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)^2$$

Polynomial degree 1



$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots$$
$$(x_n, y_n)$$

Given: a set of points on the plane

Goal: find the best line that approximates the points

Error of a line:

$$f(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)^2$$

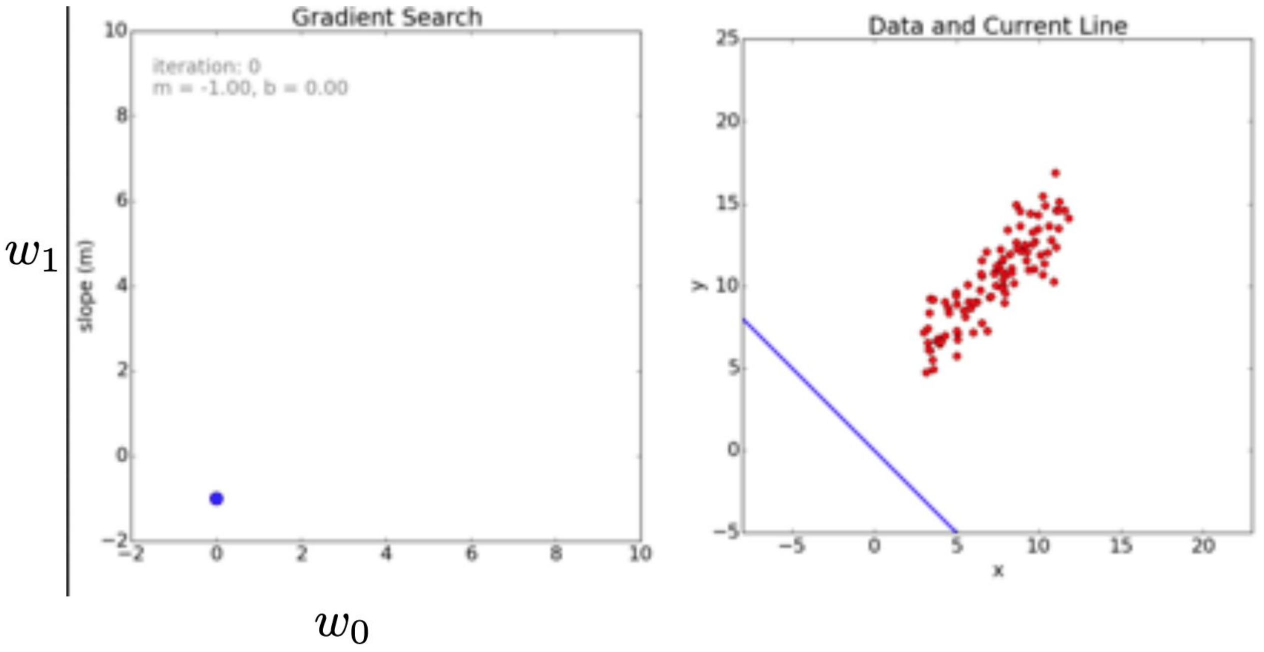
Gradient at a point:

$$\frac{\partial f(w_0, w_1)}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n} w_0 + w_1 x_i - y_i$$

$$\frac{\partial f(w_0, w_1)}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)x_i$$

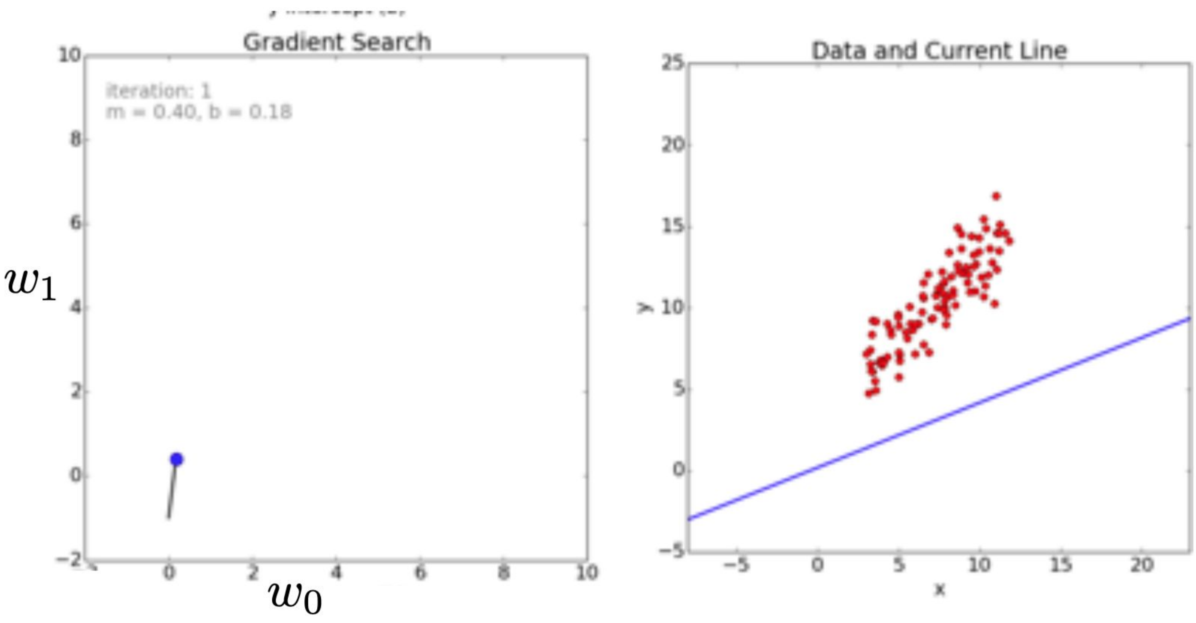
Polynomial degree 1

Iteration = 0



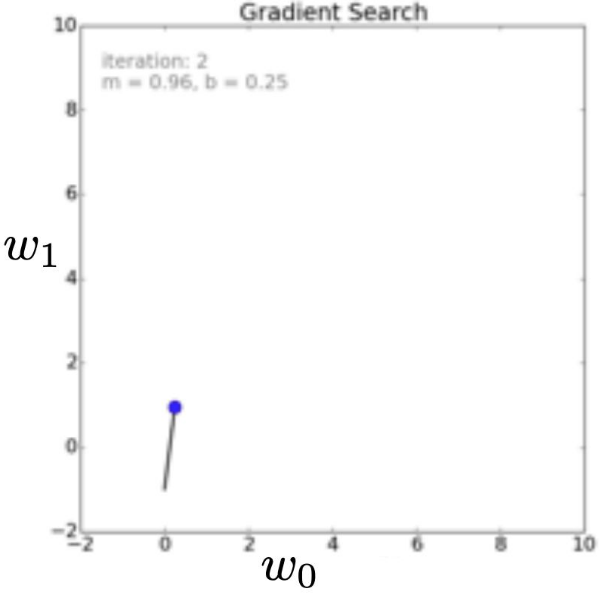
Polynomial degree 1

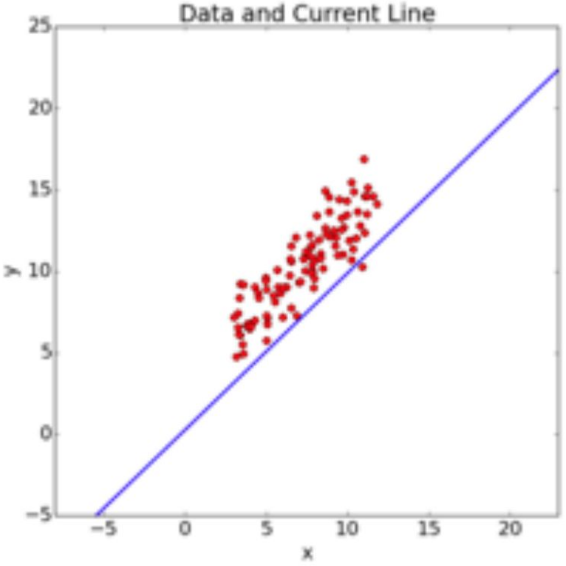
Iteration = 1



Polynomial degree 1

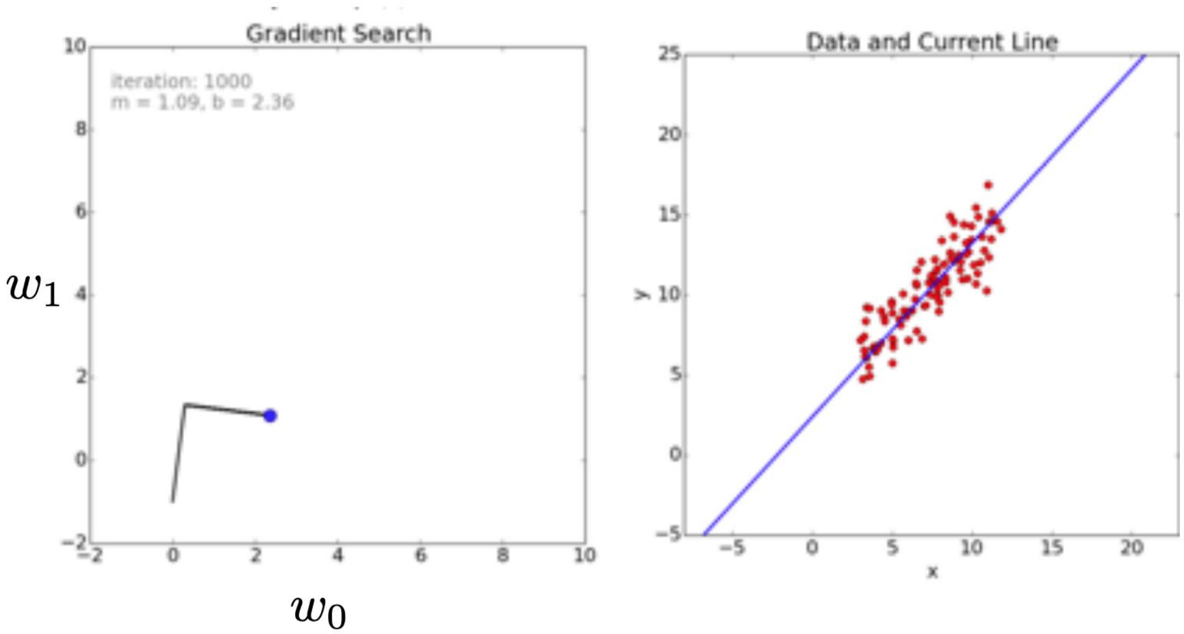
Iteration = 2





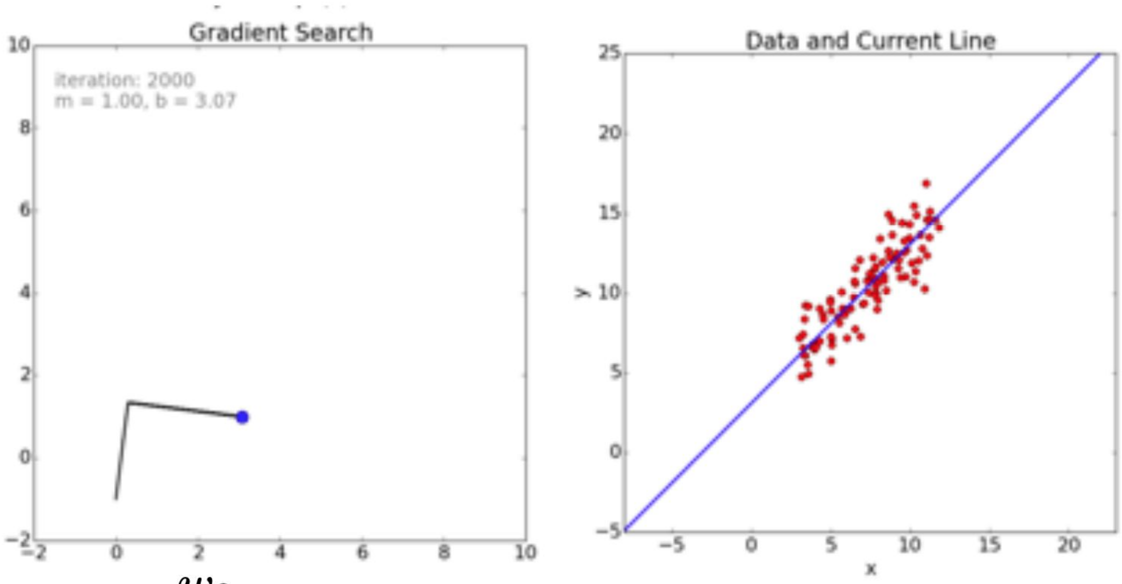
Polynomial degree 1

Iteration = 1000



Polynomial degree 1

Iteration = 2000

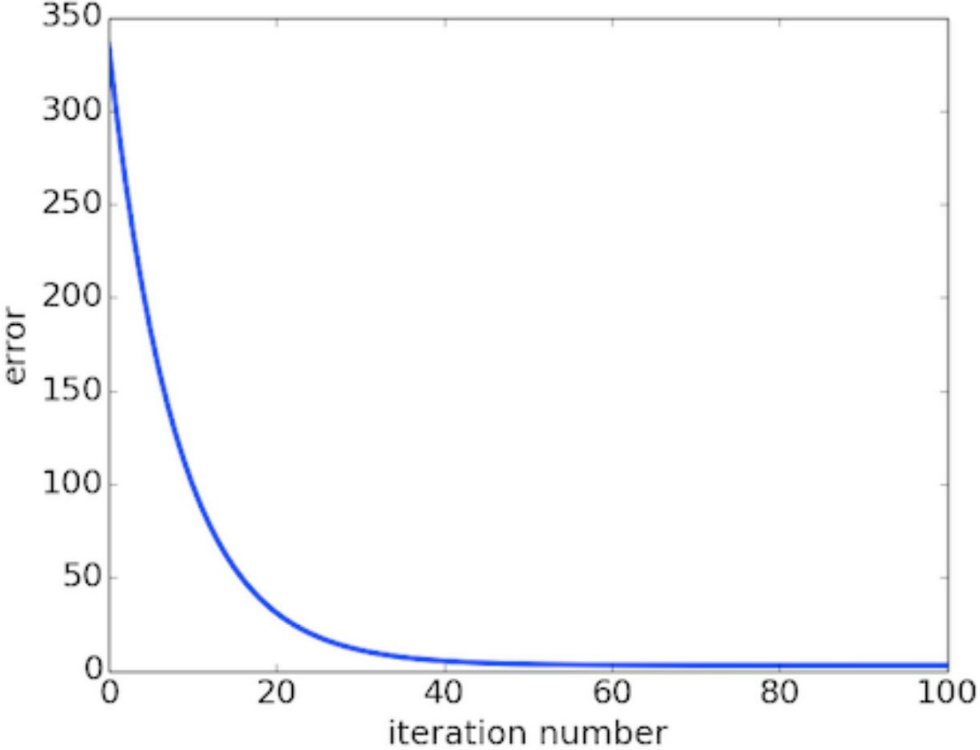


w_1

w_0

Polynomial degree 1

How error decreases



Stochastic gradient descent

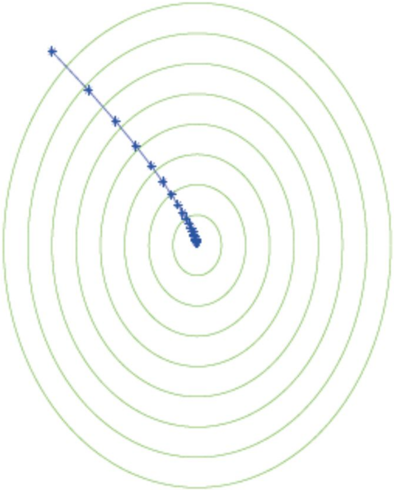
GD is not practical for large-scale data!

Consider a learning problem with millions of images?

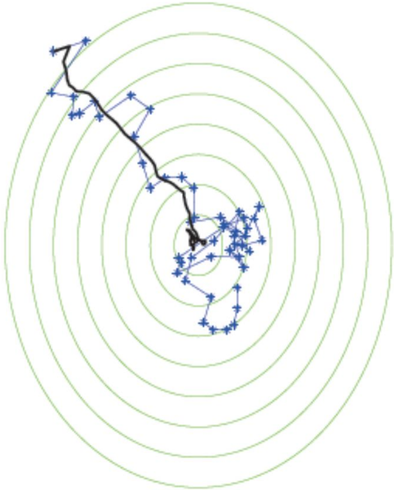
n gradient computations for n training samples per iteration!

GD versus SGD

Stochastic Gradient Descent (SGD): At each iteration, compute the gradient over a small fixed-size subset of data (min-batch)!



GD

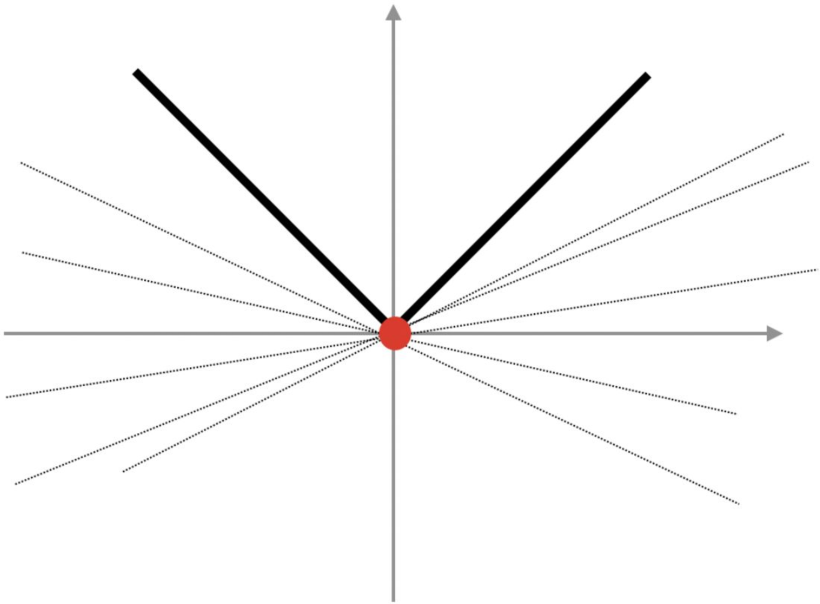


SGD

Stay tuned! We will talk about SGD in future lectures!

What if the function is not differentiable?

$$f(x) = |x|$$



To many tangent vectors at non-differentiable a point! Which direction should I take?

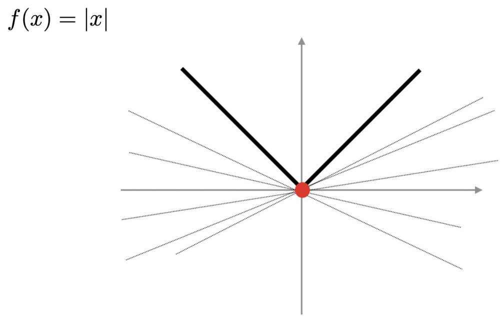
Subgradients

Definition

Let $f : \mathcal{C} \to \mathbb{R}$ be a proper function and let $\boldsymbol{x} \in \mathrm{dom}(f)$. A vector $\boldsymbol{g}$ is called a subgradient of f at $\boldsymbol{x}$ if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle \text{ for all } \boldsymbol{y} \in \mathrm{dom}(f)$$

We denote the set of all subgradients at point $\boldsymbol{x}$ by $\partial f(\boldsymbol{x})$ which is $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$ if the function is differentiable at $\boldsymbol{x}$.



Later in the course, we will introduce Subgrdient Descent algorithm!