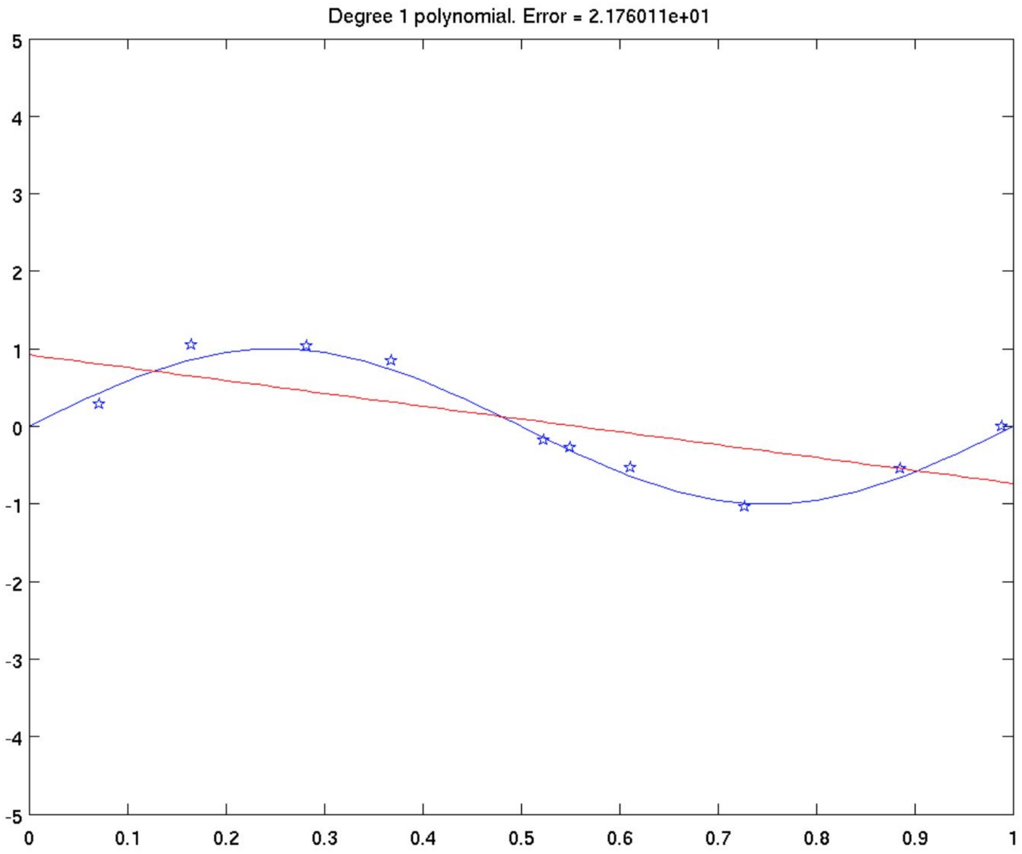
CMPSC 448: Machine Learning

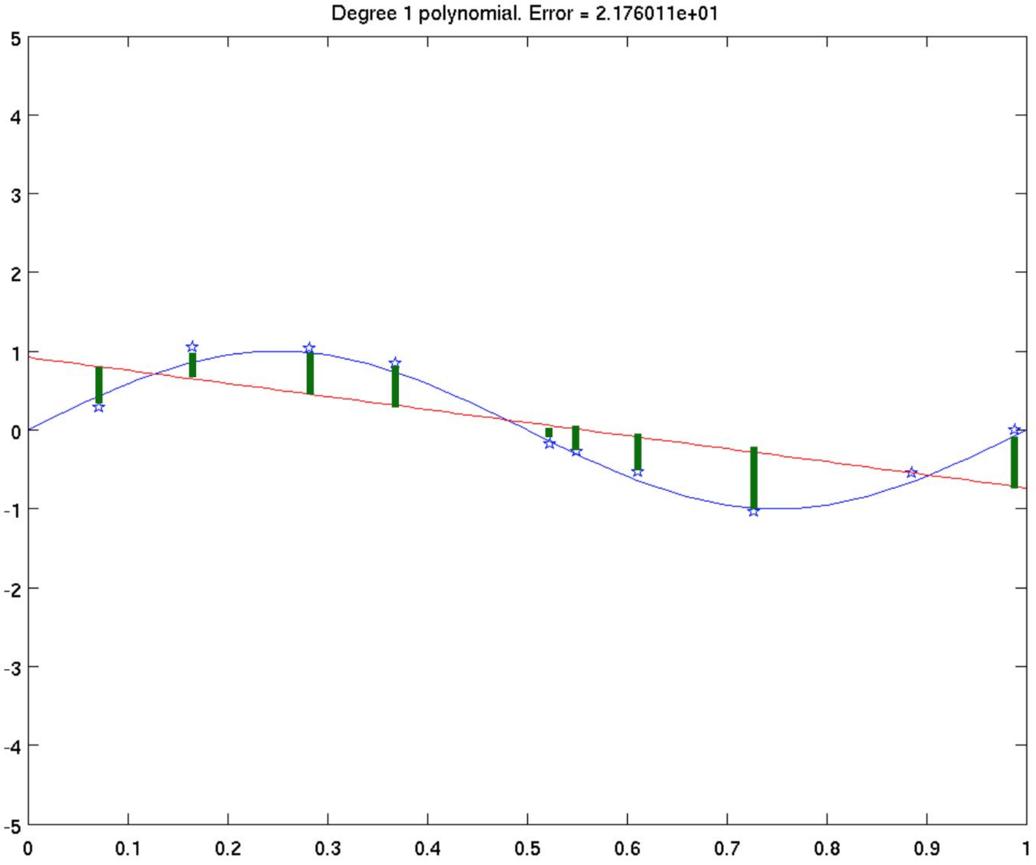
Lecture 3. Basic Convex Optimization

Rui Zhang
Fall 2022

Why optimization?



Why optimization?



Why optimization?

For a given training data:

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

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

Why optimization?

For a given training data:

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

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

Scalar valued functions of scalars

We are all familiar with basic calculus, and functions of the form $f : \mathbb{R} \to \mathbb{R}$ and their derivatives:

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}.$$

The derivative of a function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value).

- $f(x) = x^r$, then $f'(x) = rx^{r-1}$,
- $\frac{d}{dx} e^x = e^x$.
- Sums rule: $(\alpha f + \beta g)' = \alpha f' + \beta g'$ for all functions f and g and all real numbers α and β
- Product rule: $(fg)' = f'g + fg'$ for all functions f and g.
- Chain rule: If $f(x) = h(g(x))$, then

$$f'(x) = h'(g(x)) \cdot g'(x).$$

Scalar valued functions of vectors

Multivariable calculus (also known as multivariate calculus) is the extension of calculus in one variable to calculus with functions of several variables, $f : \mathbb{R}^d \to \mathbb{R}$, e.g.,

$$f(x, y) = \frac{x^2 y}{x^4 + y^2}$$

In many machine learning applications, our model can be modeled as a function f that takes d features as inputs and maps it to a real number (regression) or binary variable (classification), e.g.,

- Linear regressions: let's $\boldsymbol{x} \in \mathbb{R}^d$ be d features of a samples and $\boldsymbol{w} \in \mathbb{R}^d$ be parameter vector of of a linear model f, then the prediction is:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = \sum_{i=1}^{d} w_i x_i$$

So, we need to have a basic understanding of multivariable calculus.

Gradient (first order)

Definition

Let $f : \mathcal{C} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, the gradient of f at $\boldsymbol{x} \in \mathcal{C}$ is the vector in $\mathbb{R}^d$ denoted by $\nabla f(\boldsymbol{x})$ and defined by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\boldsymbol{x}) \end{bmatrix}$$

An an example, let's consider the function $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^d x_i y_i$, then

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial \sum_{i=1}^d x_i y_i}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{i=1}^d x_i y_i}{\partial x_d} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_d \end{bmatrix} = \boldsymbol{y}$$

Hessian (second order)

Definition

Let $f : \mathcal{C} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Then, the Hessian of f at $\boldsymbol{x} \in \mathcal{C}$ is the vector in $\mathbb{R}^d$ denoted by $\nabla^2 f(\boldsymbol{x})$ and defined by

$$\nabla^2 f(\boldsymbol{x}) = \left[\frac{\partial^2 f}{\partial x_i x_j}(\boldsymbol{x})\right]_{1\le i,j\le d} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

Hessian (second order)

Definition

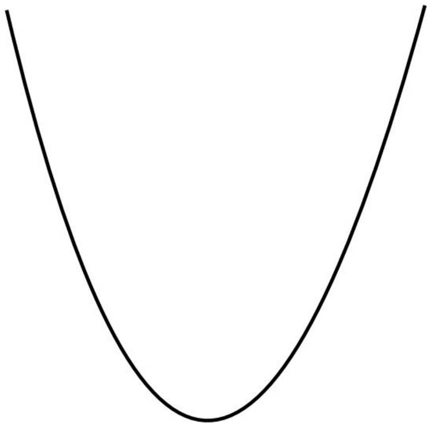
Let $f : \mathcal{C} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Then, the Hessian of f at $\boldsymbol{x} \in \mathcal{C}$ is the vector in $\mathbb{R}^d$ denoted by $\nabla^2 f(\boldsymbol{x})$ and defined by

$$\nabla^2 f(\boldsymbol{x}) = \left[\frac{\partial^2 f}{\partial x_i x_j}(\boldsymbol{x})\right]_{1 \le i,j \le d} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

$$f(x, y) = x^3 - 2xy - y^6 \xrightarrow{\textbf{Hessian}} \begin{bmatrix} 6x & -2 \\ -2 & -30y^4 \end{bmatrix}$$

One minute calculus

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

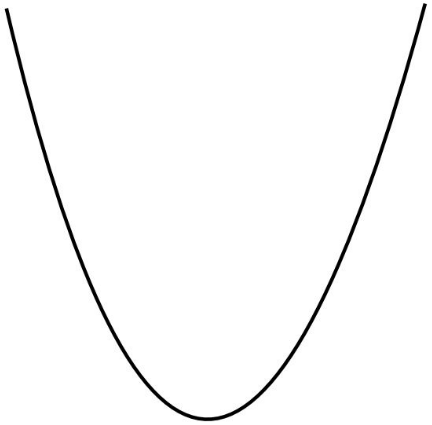


$$f(x) = (x-2)^2$$

One minute calculus

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

Easy: set the derivative to zero!



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

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

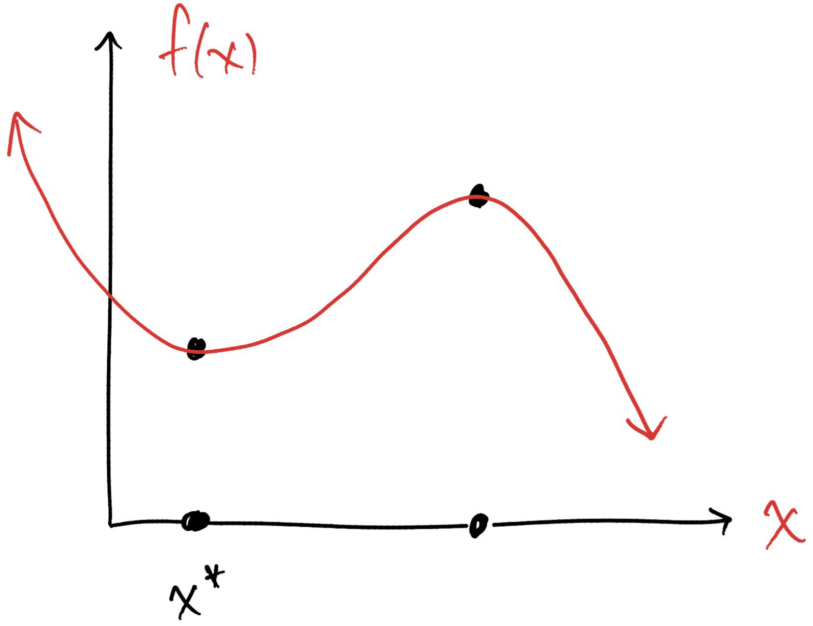
$$\Rightarrow \quad x_* = 2$$

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

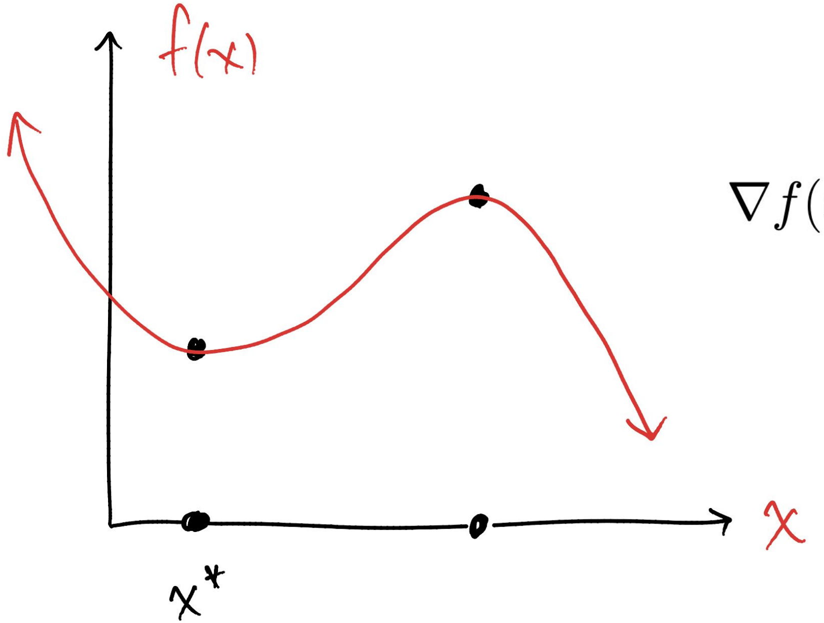
Property 1

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?

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

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

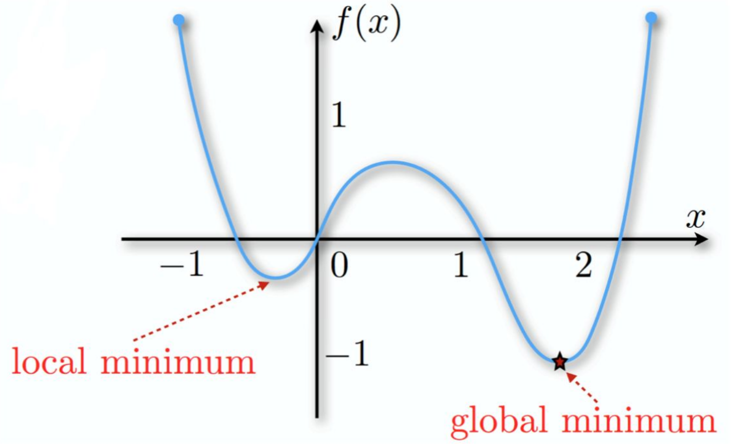
(1) There might not be a closed form solution for $f'(x) = 0$!

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., $z^\intercal \nabla^2 f(\boldsymbol{x}^*) z \geq 0, \quad \forall z \in \mathbb{R}^d$).

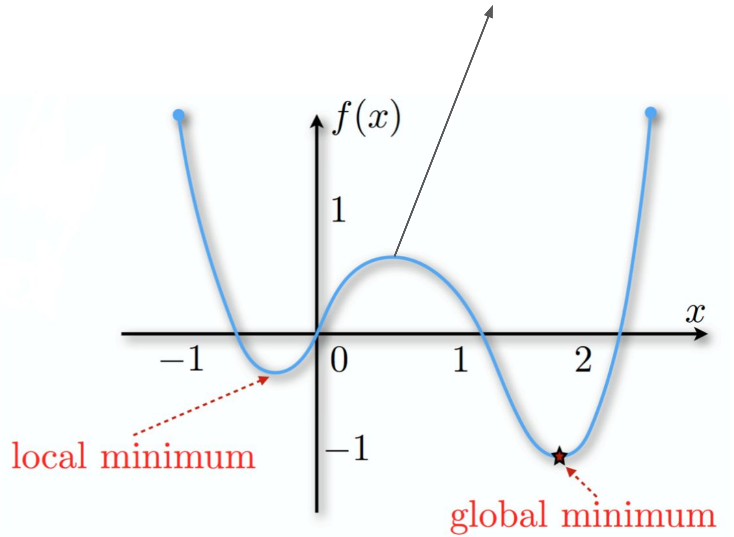
Property 2

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

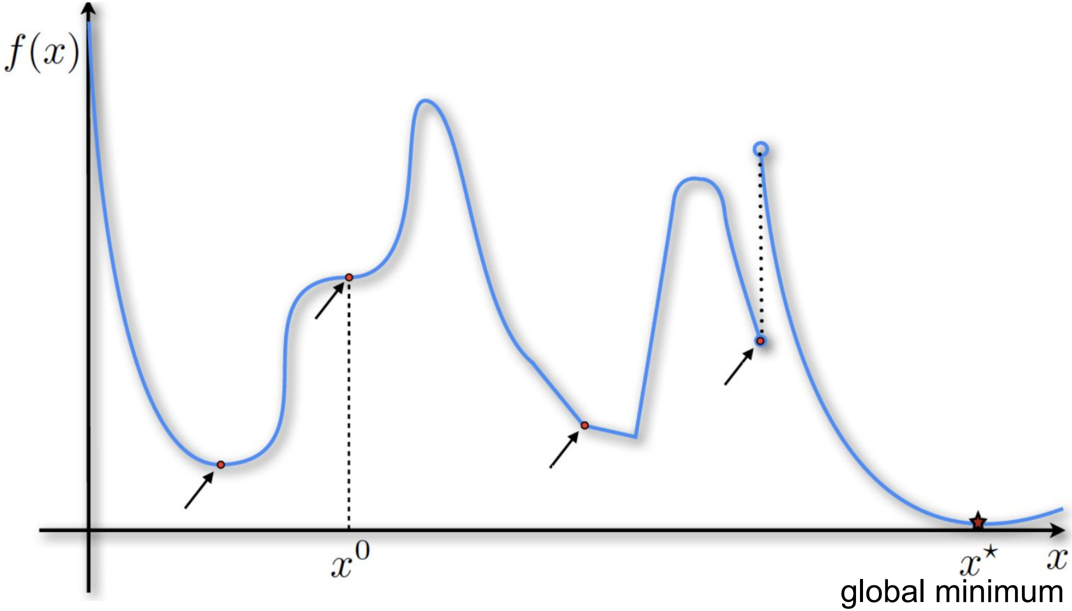
$$\min_{x \in \mathbb{R}} \quad x^4 - 3x^3 + x^2 + \frac{3}{2}x$$

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



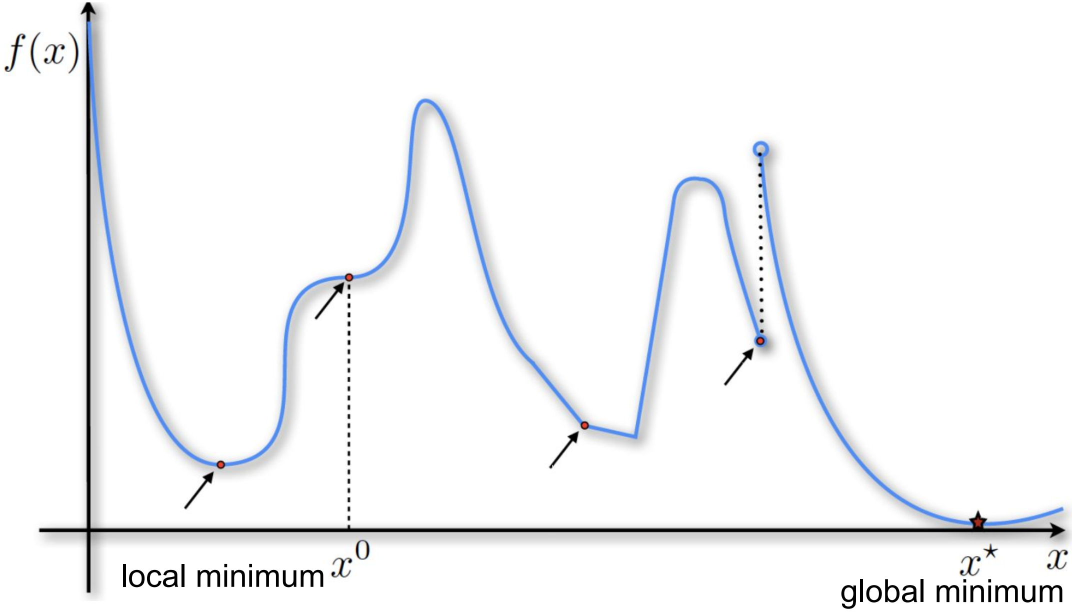
Iterative optimization

Fog of war



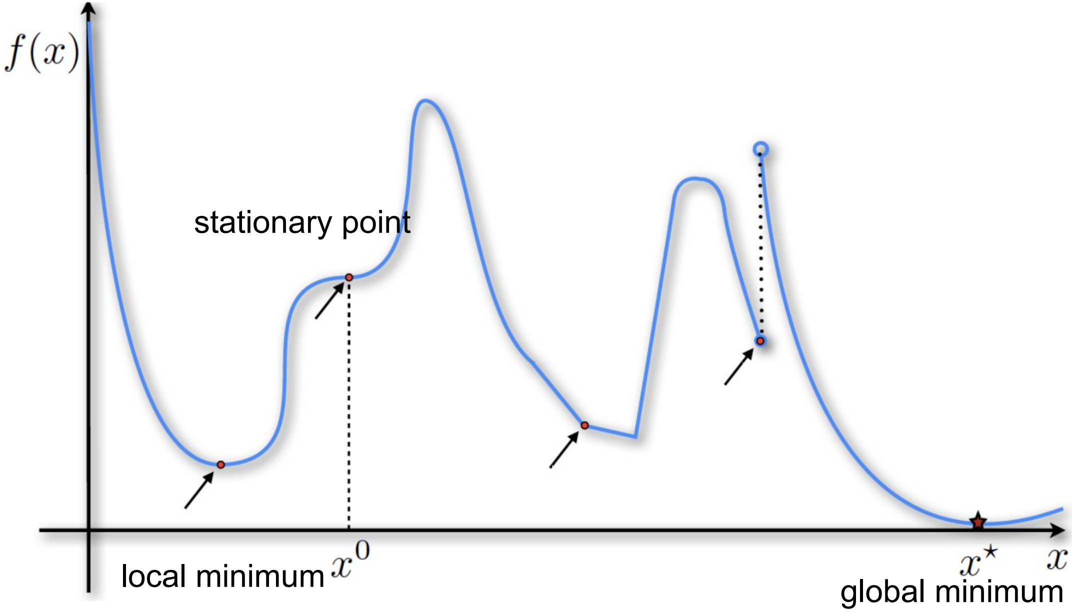
Iterative optimization

Fog of war



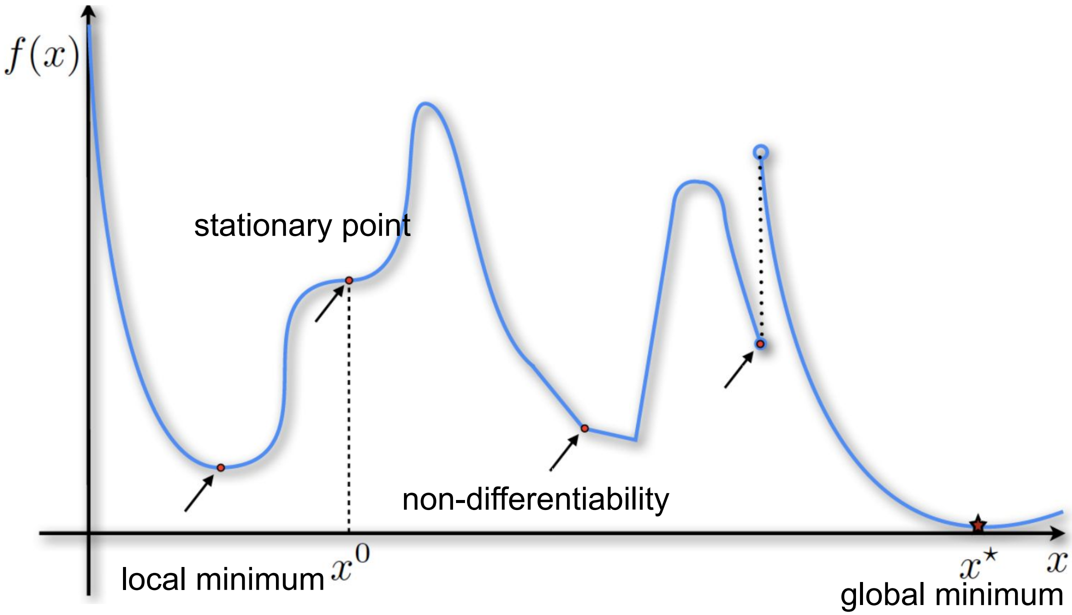
Iterative optimization

Fog of war



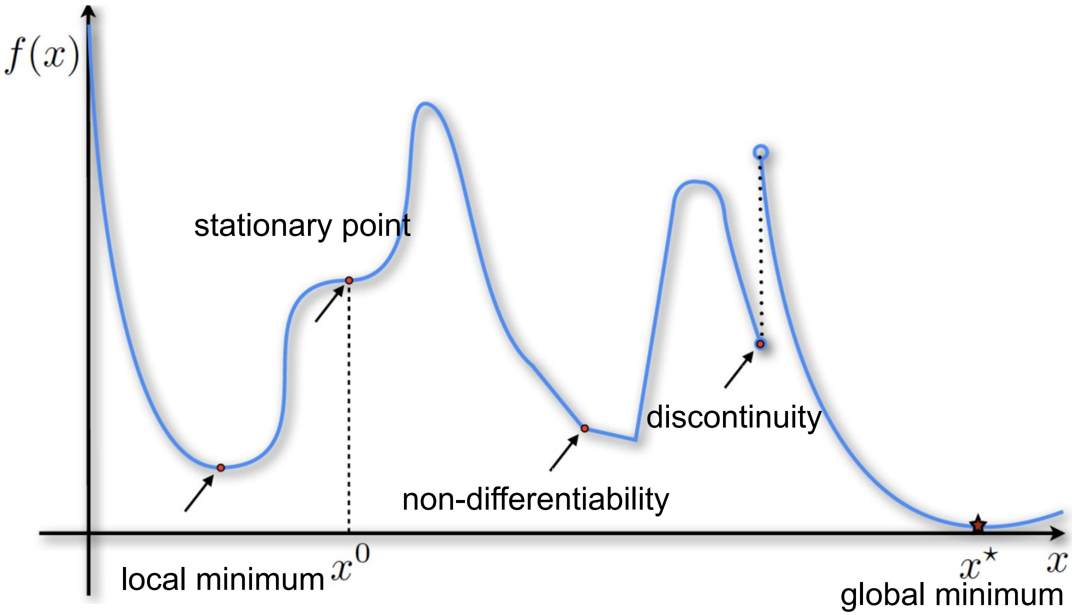
Iterative optimization

Fog of war



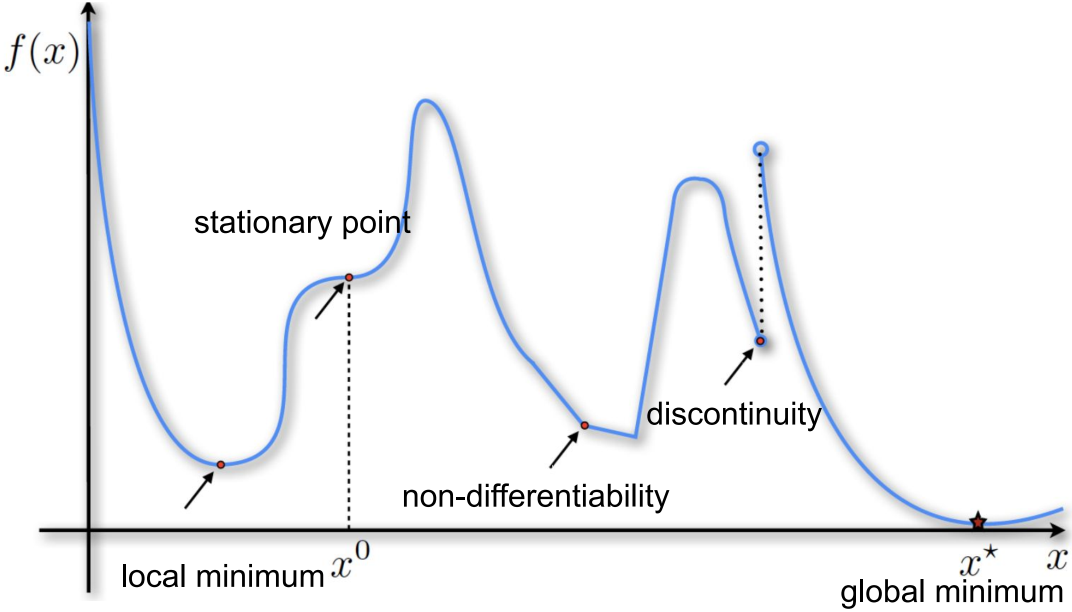
Iterative optimization

Fog of war



Iterative optimization

Fog of war



We need a key structure on the function: 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:

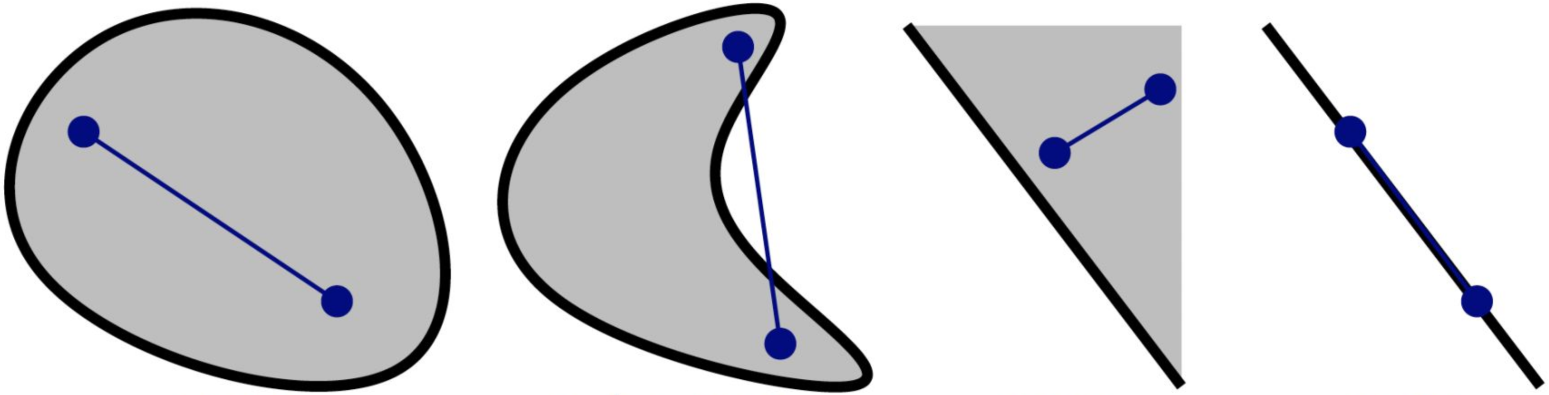
$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}$$

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:

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

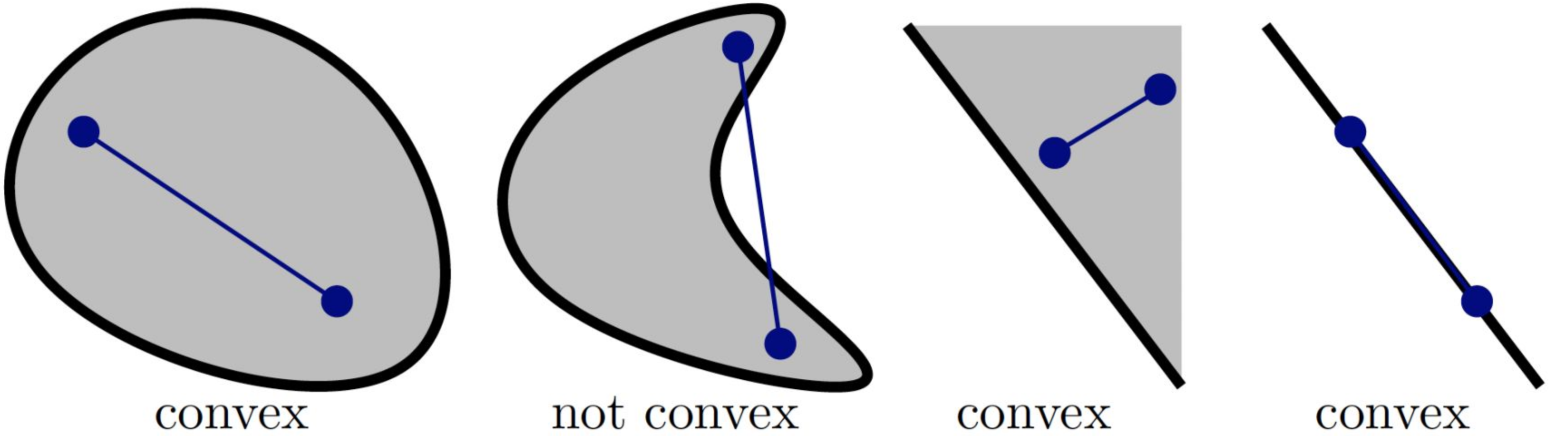


Convex set

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:

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

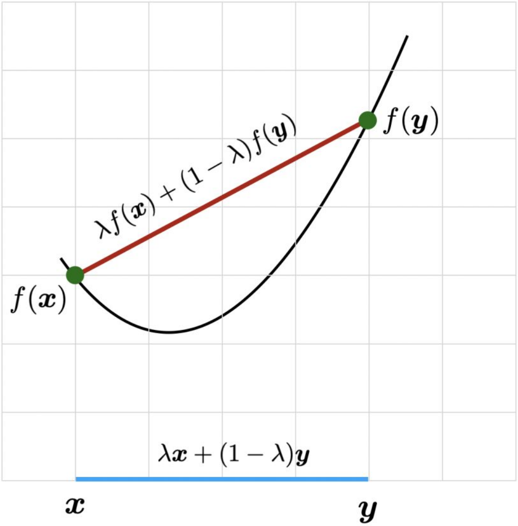


Convex function

Definition

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

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

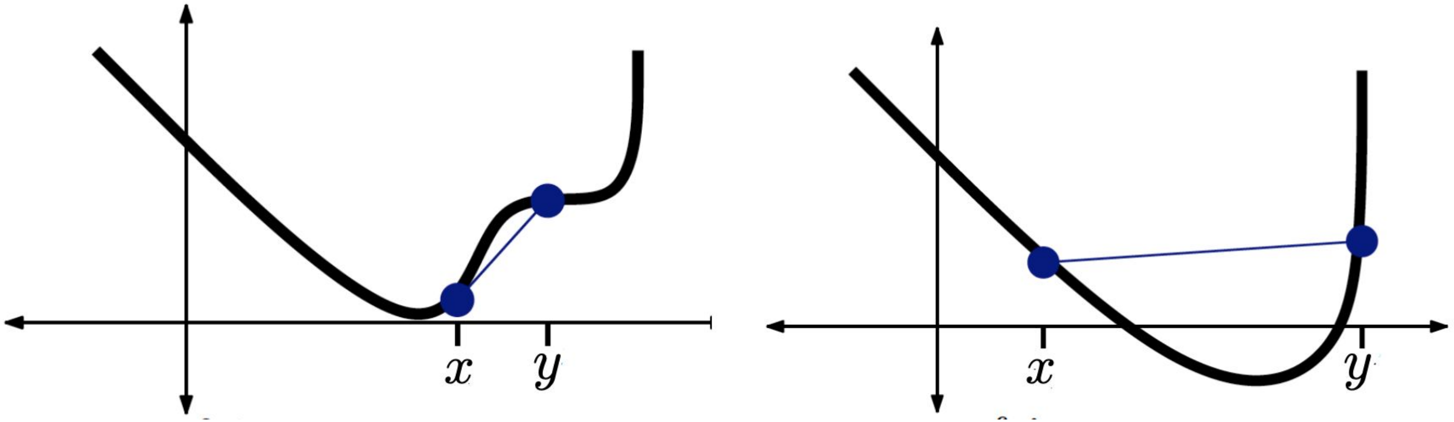


Convex function

Definition

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

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

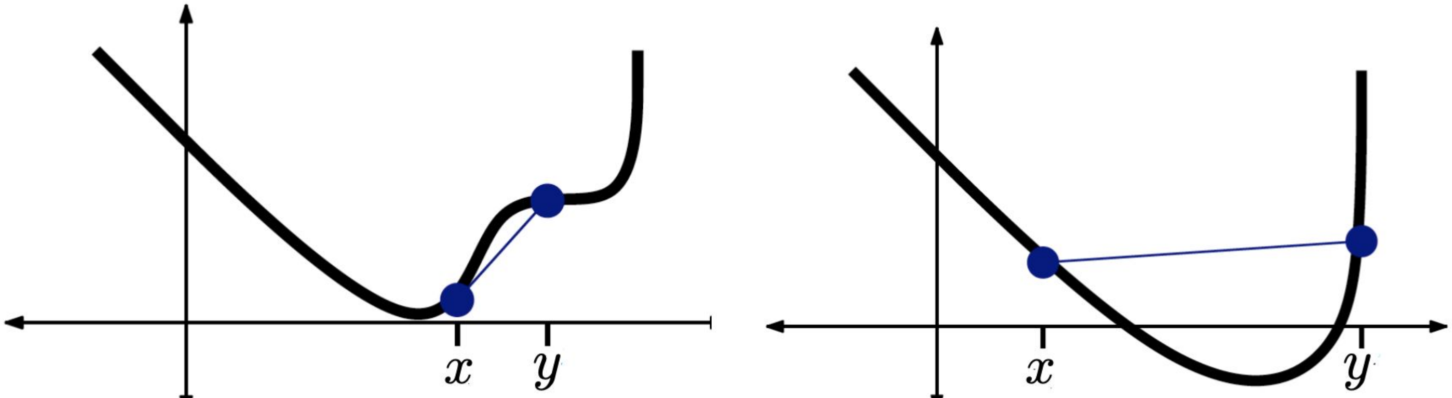


Convex function

Definition

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

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



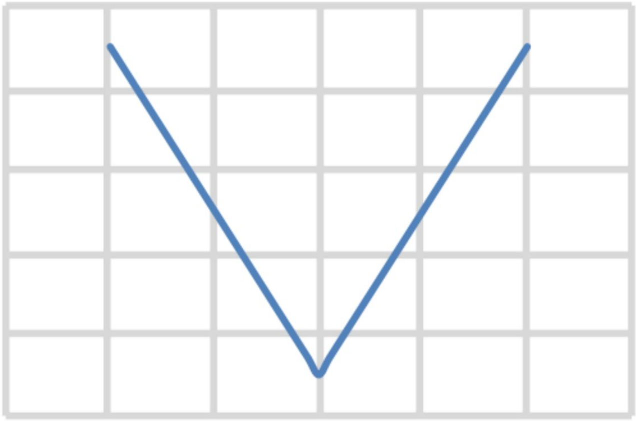
f is not convex

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 $\boldsymbol{x} \in \mathbb{R}^d$?

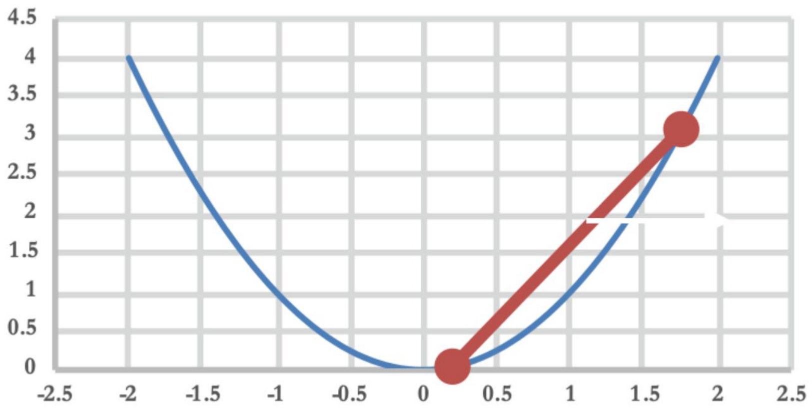
Example: norm

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

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

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

Example: quadratic

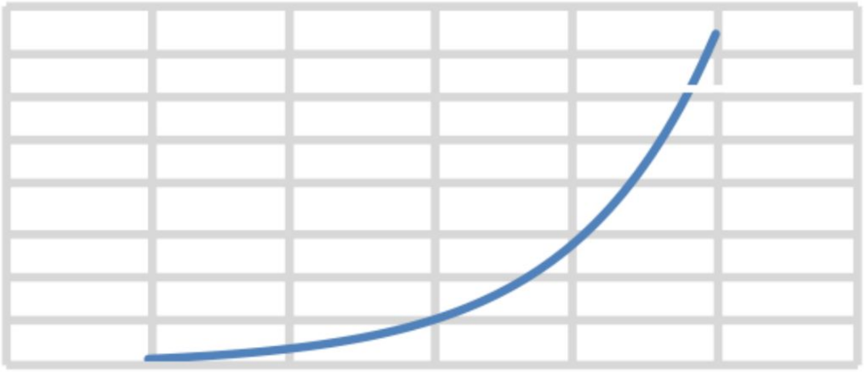


$f(x) = x^2$

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

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:

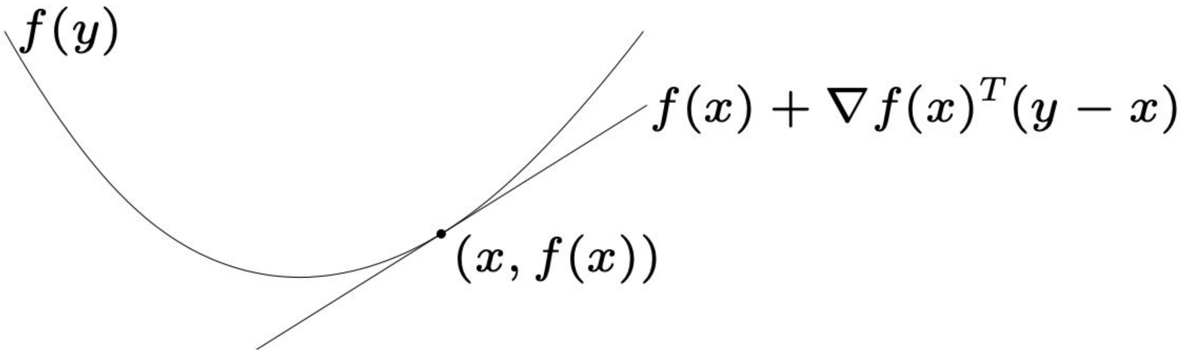
$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathtt{dom}(f), \quad f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

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 ?

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

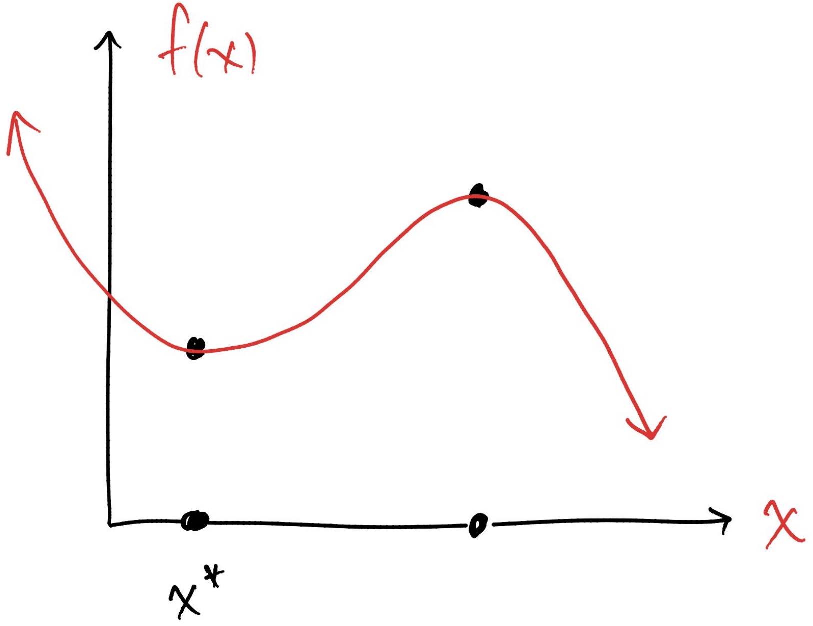
Yes, it is!

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!

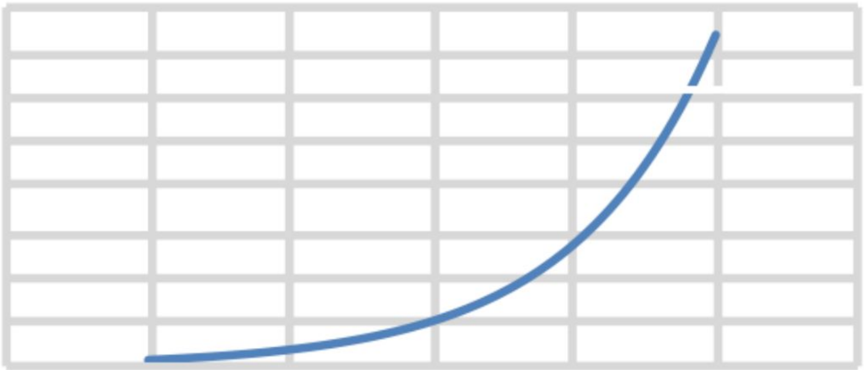
Property 6: twice continuously differentiable functions

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

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

- If the function $f : \mathbb{R}^d \to \mathbb{R}$ 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

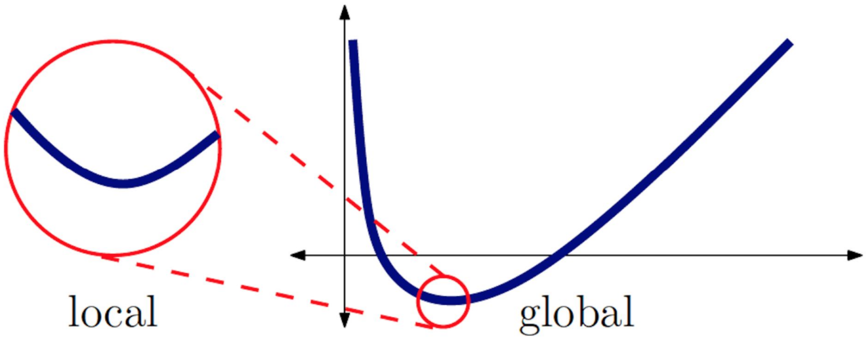
Convex Optimization

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!



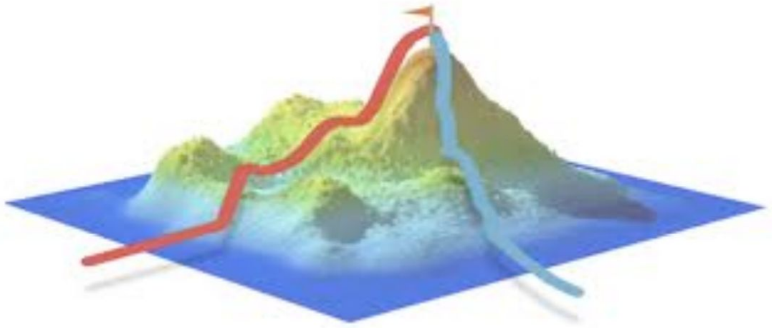
Convex functions are EASY to solve!

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 $\boldsymbol{x}_{t+1}$

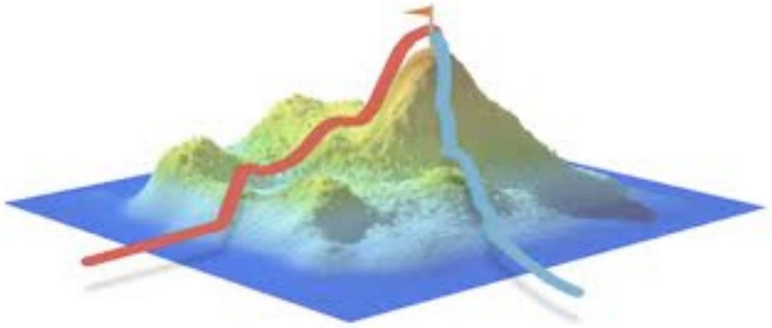


such that

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

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)

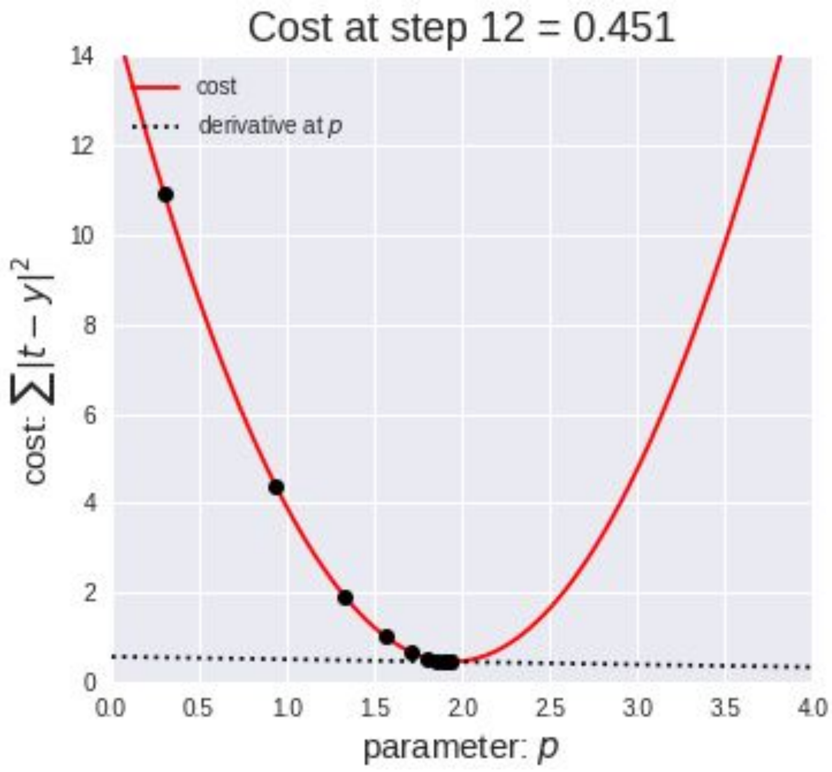
Initialize $\boldsymbol{x}_0$

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

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

Step size

Example



Cost at step 12 = 0.451
cost
derivative at p
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
cost: $\sum |t-y|^2$
parameter: p

Step size selection

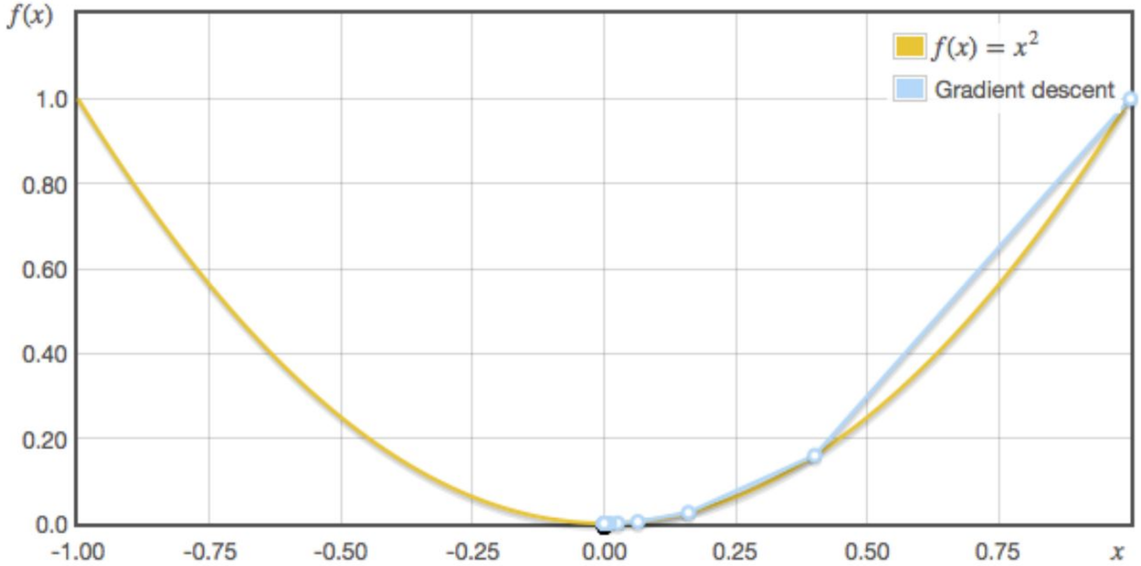
How do I choose the step size?

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

Example Step size selection

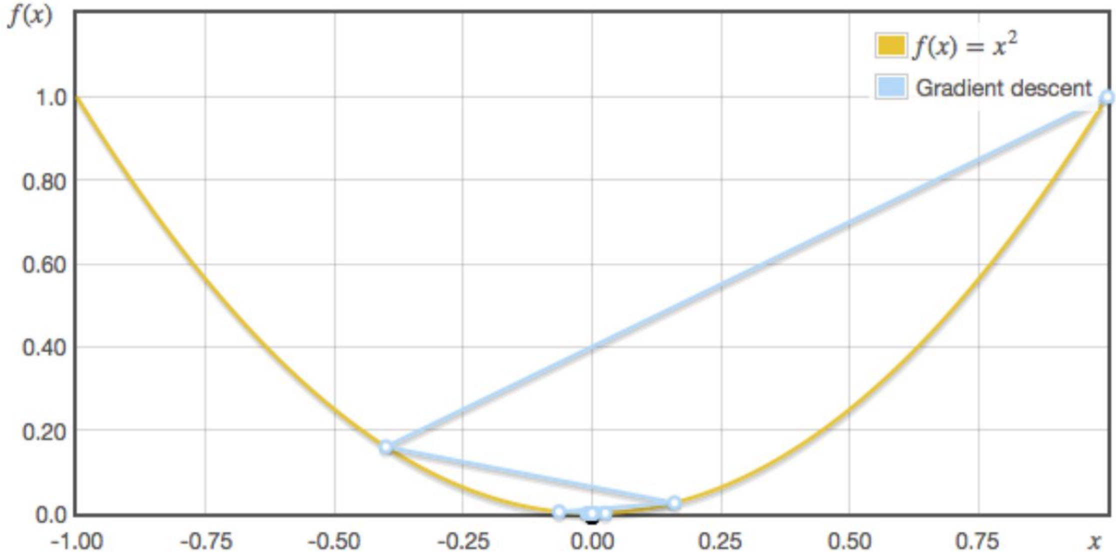
$$f(x) = x^2$$

$$\eta = 0.3$$



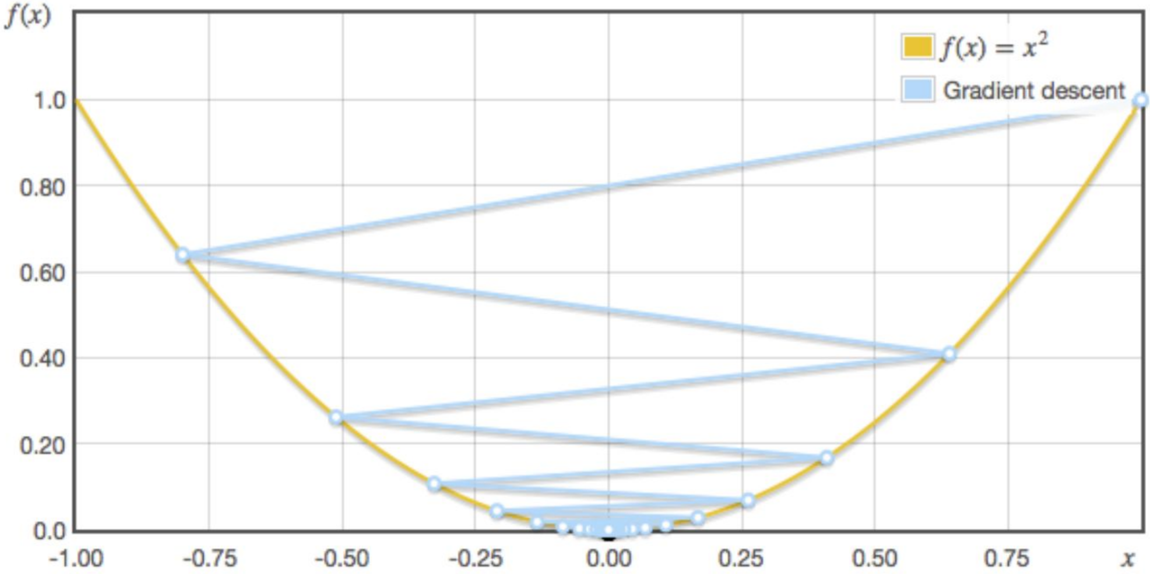
Example Step size selection

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



Example Step size selection

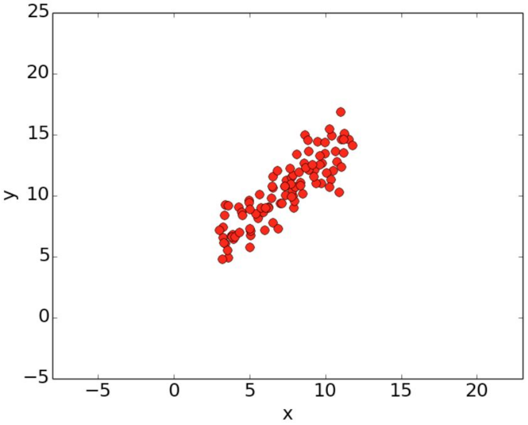
$$f(x) = x^2$$
$$\eta = 0.9$$



Polynomial degree 1

Given: a set of points on the plane

Goal: find the best line that approximates the points

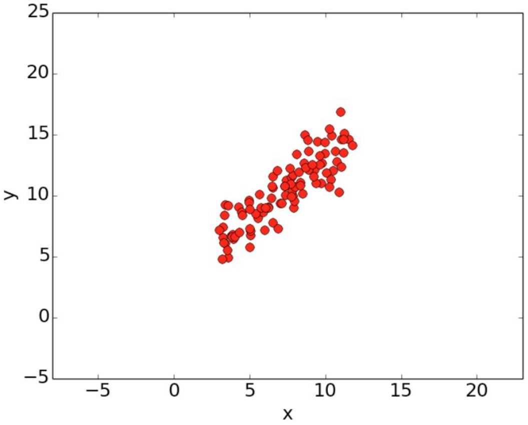


$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots$$
$$(x_n, y_n)$$

Polynomial degree 1

Given: a set of points on the plane

Goal: find the best line that approximates the points



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

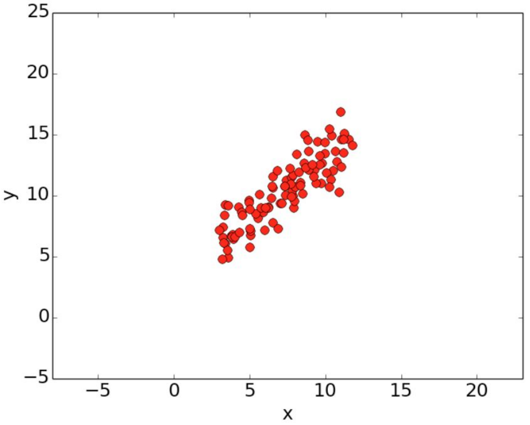
Error of a line:

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

Polynomial degree 1

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:

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

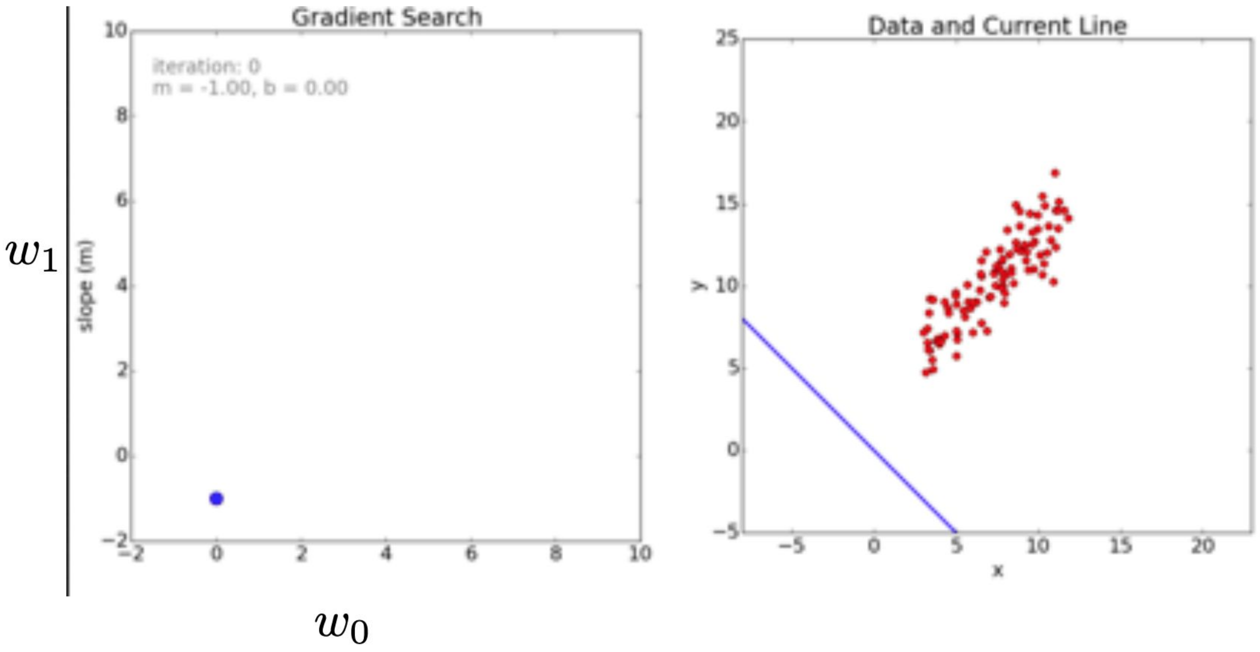
Gradient at a point:

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

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

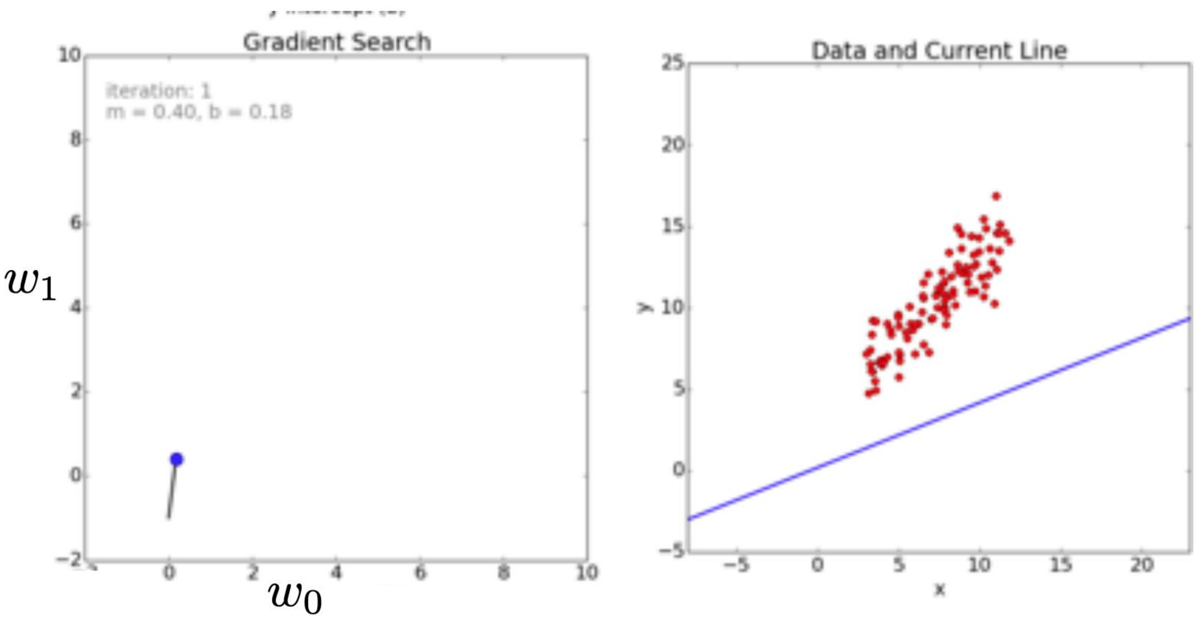
Polynomial degree 1

Iteration = 0



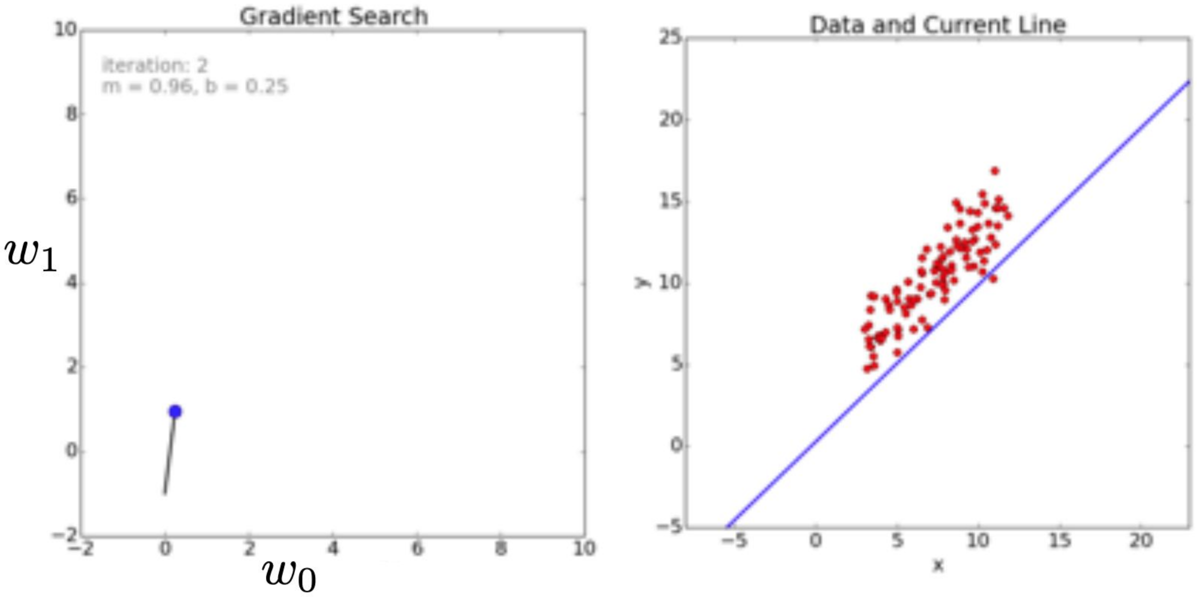
Polynomial degree 1

Iteration = 1



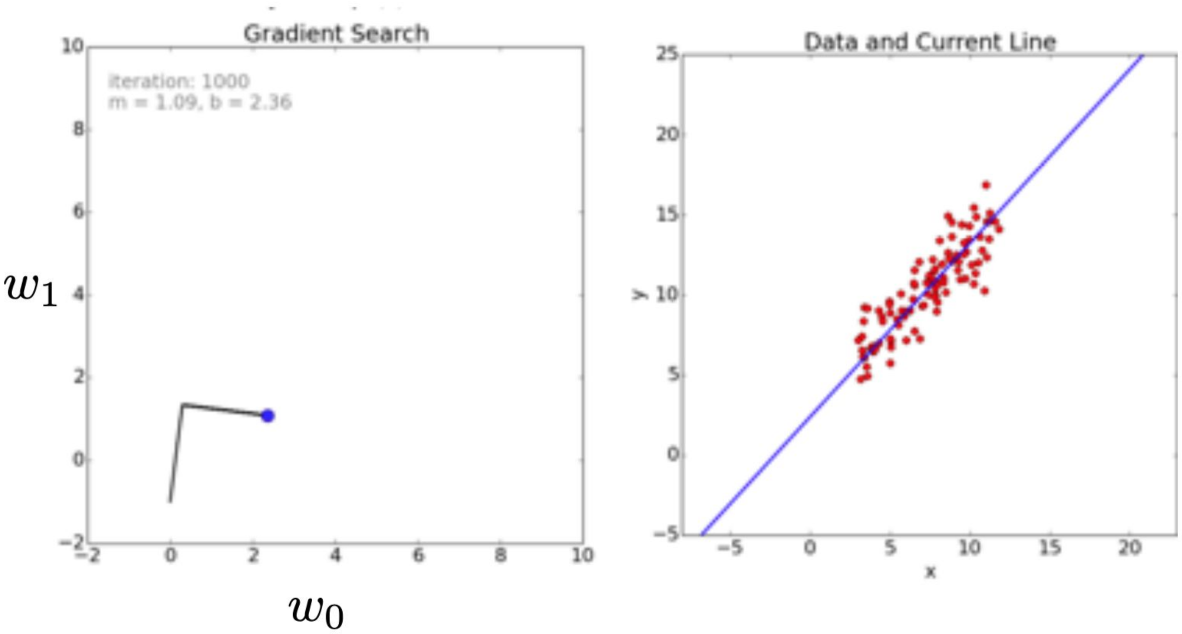
Polynomial degree 1

Iteration = 2



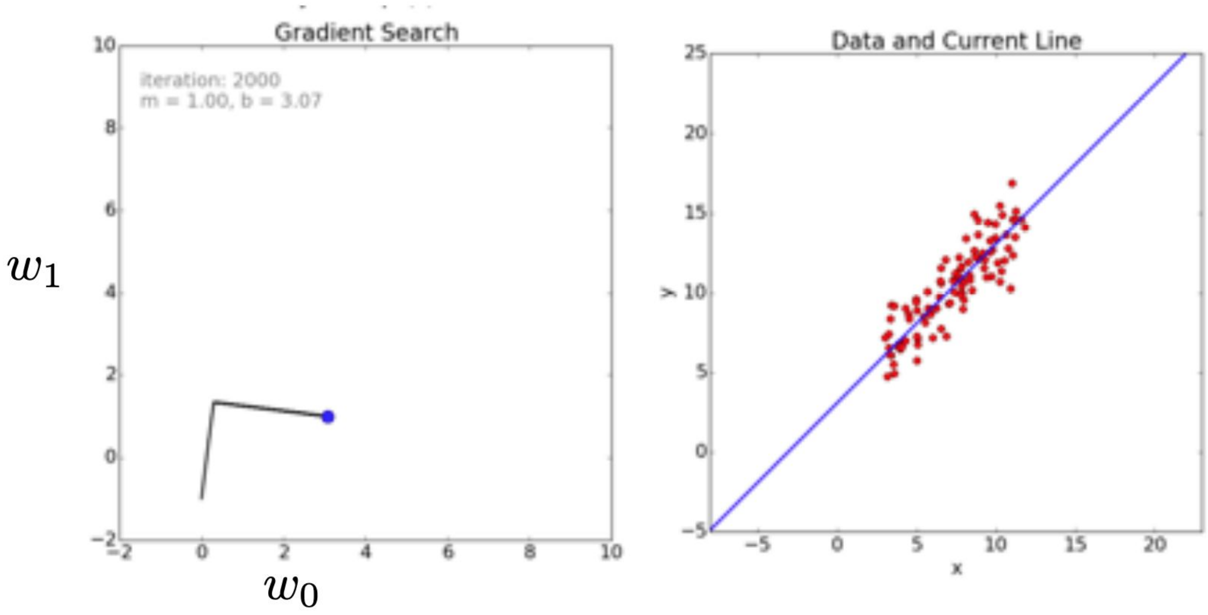
Polynomial degree 1

Iteration = 1000



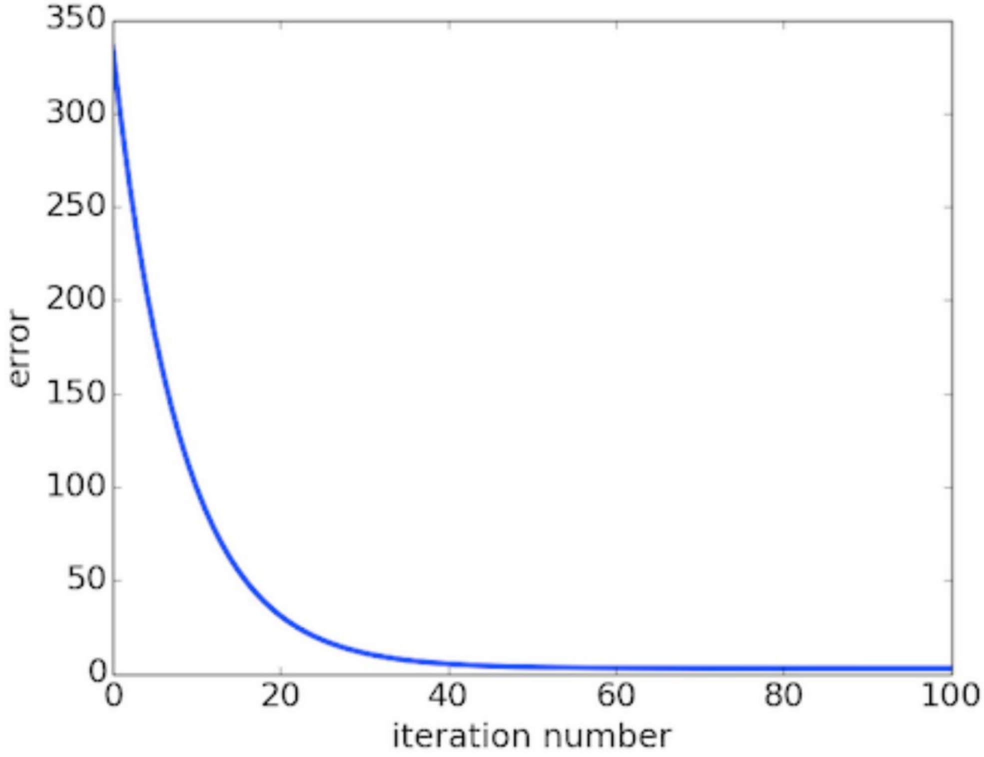
Polynomial degree 1

Iteration = 2000



Polynomial degree 1

How error decreases



Stochastic gradient descent

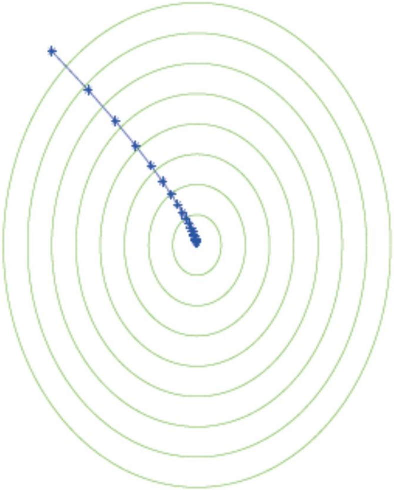
GD is not practical for large-scale data!

Consider a learning problem with millions of images?

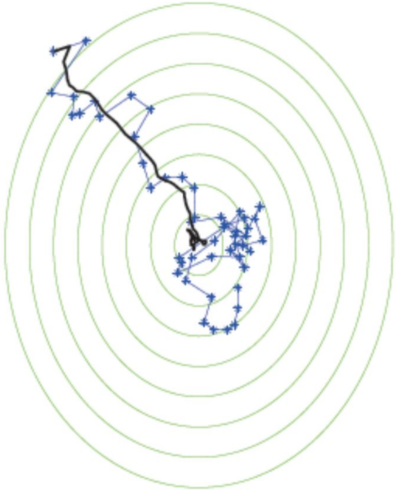
n gradient computations for n training samples per iteration!

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!