CONVEX OPTIMIZATION

Guy Azran & Itay Evron

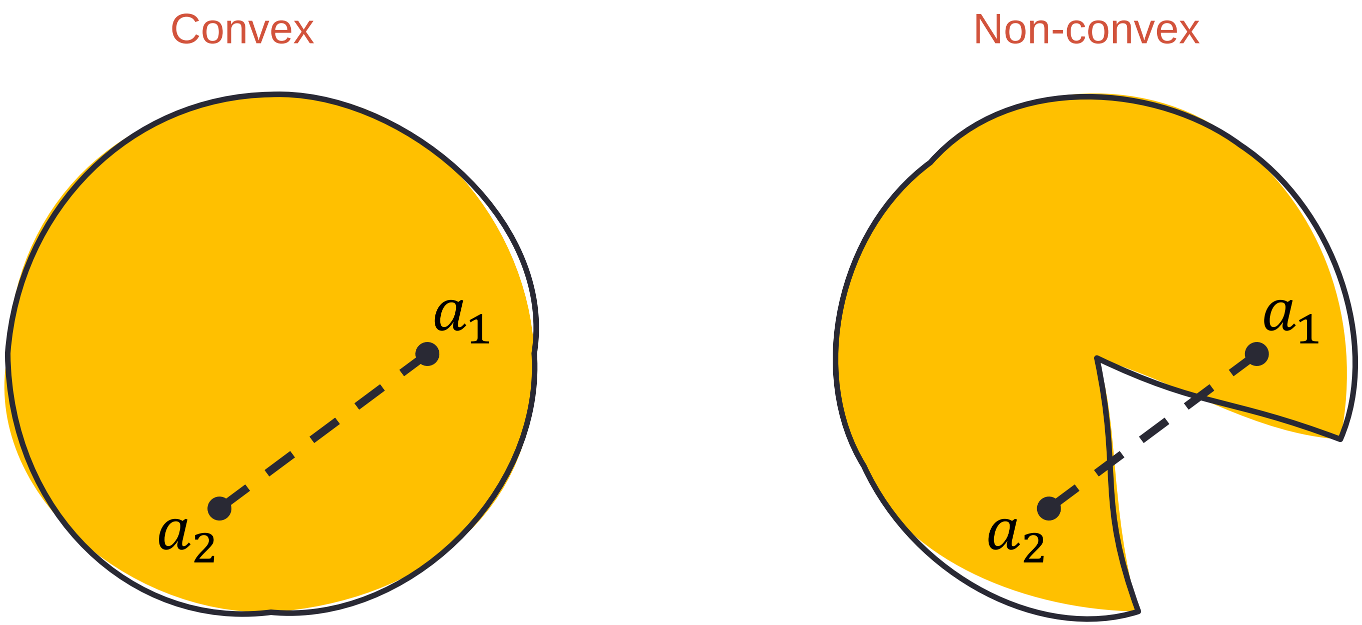
Partially based on Understanding Machine Learning by Shai Shalev-Shwartz and Shai Ben-David

Outline

- Convexity recap
- Proving convexity
- Gradient descent

Recap: Convex sets

- Intuition: C is a convex set if the line between any two points is in C.

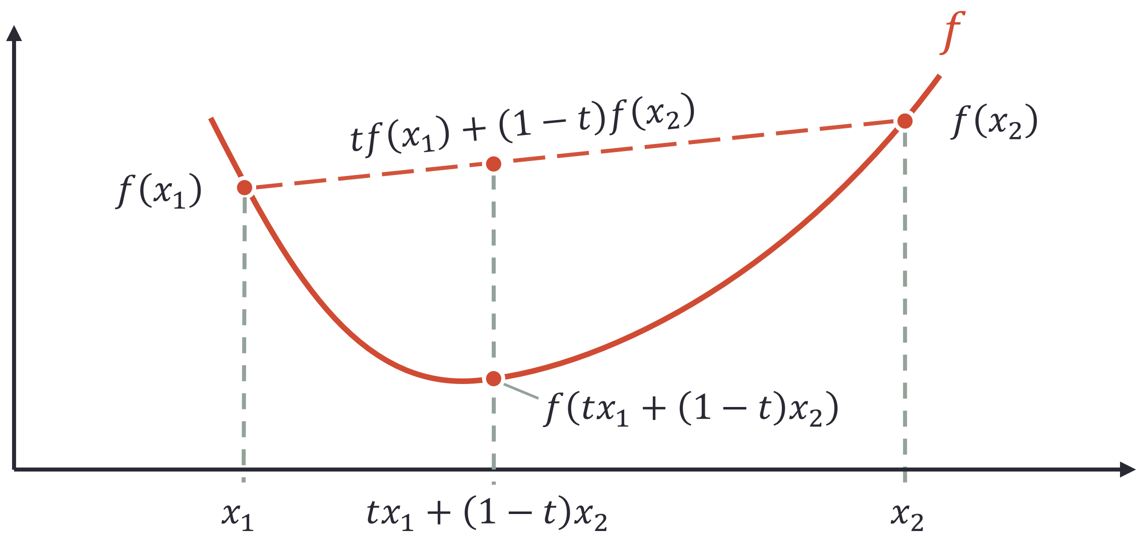


- Formally: Set $C \subset \mathcal{V}$, where $\mathcal{V}$ is some vector space, is a convex set if

$$\forall a_1, a_2 \in C\,, \forall t \in [0,1]: \quad t \cdot a_1 + (1-t) \cdot a_2 \in C$$

Recap: Convex functions

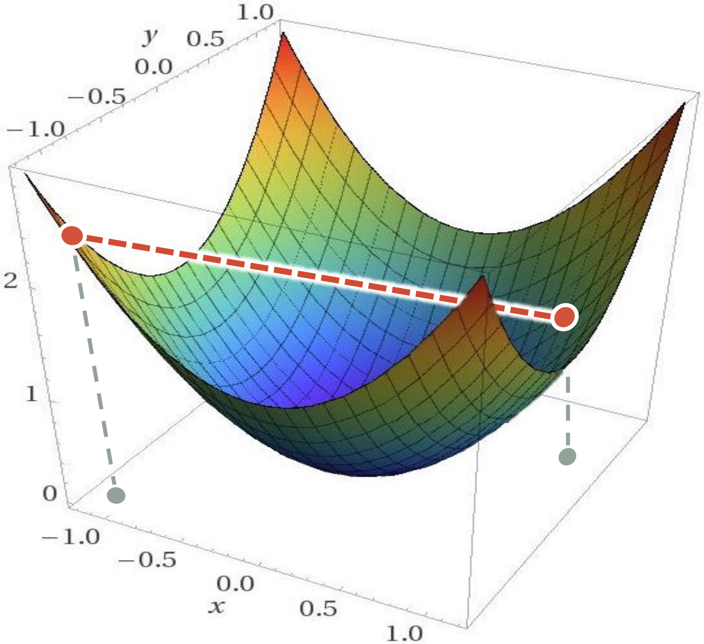
- Let C be a convex set.
- Definition: A function $f: C \to \mathbb{R}$ is a convex function if:
  - Intuitively: the line between any two points on its graph lies above the graph.



  - Formally: $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in C,\ \forall t \in [0,1]$: $tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2) \geq f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2)$

Recap: Convex functions (2d)

- Let C be a convex set.
- Definition: A function $f: C \to \mathbb{R}$ is a convex function if:
  - Intuitively: the line between any two points on its graph lies above the graph.



Source: VProexpert

Exercise: Sum of convex functions

- Prove: If $g, h: C \to \mathbb{R}$ are convex functions, then $g + h$ is convex.
- Proof:
  - Let $t \in [0,1]$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$.
  - Then

$$t(g+h)(\boldsymbol{x}_1) + (1-t)(g+h)(\boldsymbol{x}_2)$$

$$= tg(\boldsymbol{x}_1) + th(\boldsymbol{x}_2) + (1-t)g(\boldsymbol{x}_2) + (1-t)h(\boldsymbol{x}_2)$$

$$= \underbrace{tg(\boldsymbol{x}_1) + (1-t)g(\boldsymbol{x}_2)} + \underbrace{th(\boldsymbol{x}_1) + (1-t)h(\boldsymbol{x}_2)}$$

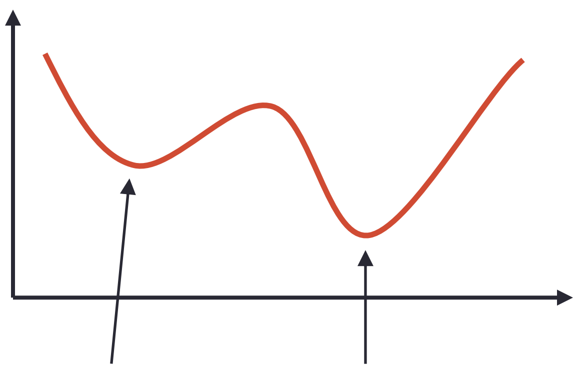
$$\geq g(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) + h(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2)$$

- Extra: generalize the above to a finite sum of arbitrary size

Convexity: Motivation

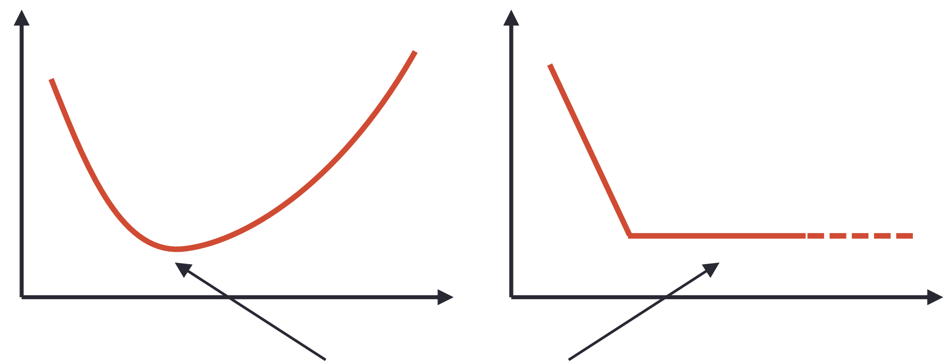
- Theorem: Any local minimum of a convex function is a global minimum.
  - Note: there may be more than one minimizer.

Non-convex function



local, not global local and global

Convex functions



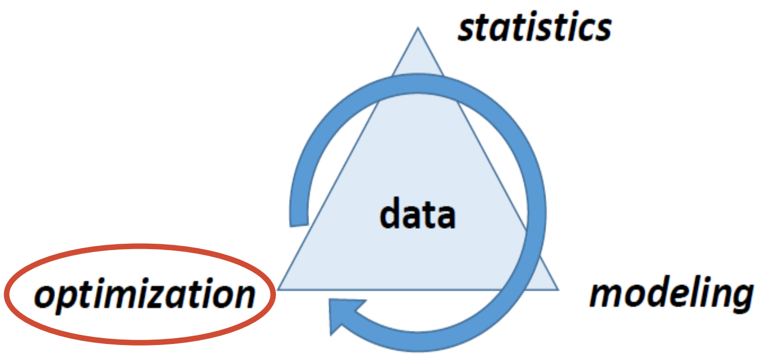
any local minimum is global

Convexity: Motivation

- Theorem: Any local minimum of a convex function is a global minimum.

- **Why is this interesting?**

  Because… Optimization!



Convex landscape (illustration)

Source: Wikipedia

PROVING CONVEXITY

Mathematical tools for efficiently testing convexity

A shortcut to convexity: the Hessian

- Remember the Hessian matrix? $$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$
- Turns out it can help prove the convexity of a function!
- Theorem: a twice-differentiable function $f: C \to \mathbb{R}$ is convex <u>iff</u> $\nabla^2 f \succcurlyeq 0$
- Example: when is a 1d parabola $f(x) = ax^2 + bx + c$ convex?
- Answer: if and only if $\nabla^2 f = \left[\frac{\partial^2 f}{\partial x^2}\right] = [2a] \succcurlyeq 0 \Leftrightarrow a \geq 0$

Retrospect: Hard-SVM is convex

- Recall the Hard-SVM problem formulation:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{w}\|_2^2$$

$$\text{s.t.} \quad y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1, \quad \forall i \in [m]$$

Denote $f(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2$, the minimized objective

Equivalent to constraining the minimized $\boldsymbol{w}$ to:

$$C = \{\boldsymbol{w} \in \mathbb{R}^d \mid \forall i \in [m]: \ y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1\}$$

- Equivalent formulation:

$$\underset{\boldsymbol{w}\in C}{\operatorname{argmin}} \ f(\boldsymbol{w})$$

- We wish to show that the Hard-SVM problem is convex.
  - We will show that the objective is convex, and then that C is a convex set.

The objective is convex

- Recall: a twice-differentiable function $f: C \to \mathbb{R}$ is convex iff $\nabla^2 f \succcurlyeq 0$

- Exercise: prove that $f(\boldsymbol{w}) = \|\boldsymbol{w}\|^2$ is convex
  - Detour: how would we solve this without the Hessian?

$$\frac{\partial}{\partial w_i \partial w_j} \|\boldsymbol{w}\|_2^2 = \frac{\partial}{\partial w_i \partial w_j} \left(\sum_k {w_k}^2 \right) = \frac{\partial}{\partial w_i} 2 w_j = \begin{cases} 2, & i = j \\ 0, & i \neq j \end{cases}$$

$$\Longrightarrow \nabla^2 \|\boldsymbol{w}\|_2^2 = 2 \mathrm{I}_d \succ 0$$

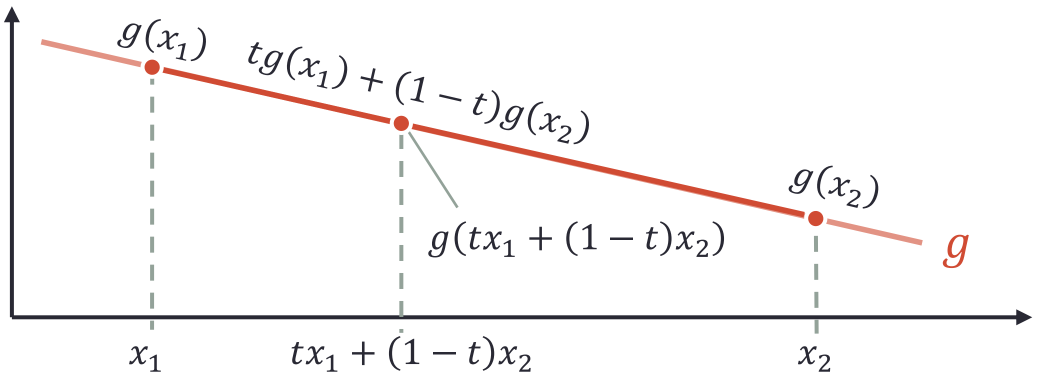
- Voilà! $\|\boldsymbol{w}\|_2^2$ is convex!

Other properties of convexity

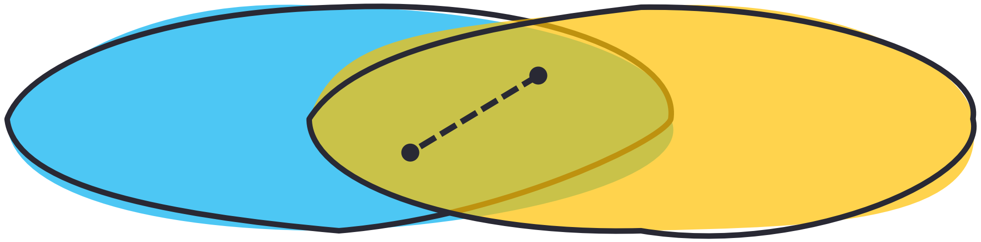
- Lemma 1: Any linear function $g(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b$ is convex and holds

$$g(t \cdot \boldsymbol{x} + (1-t)\boldsymbol{x}) = t \cdot g(\boldsymbol{x}) + (1-t)g(\boldsymbol{x})$$

in equality



- Lemma 2: Any intersection of convex sets is a convex set.



The constraints are convex

- Claim: the set $C = \{\boldsymbol{w} \in \mathbb{R}^d \mid \boxed{\forall i \in [m]:}\ y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1\}$ is convex.
- Proof:

  $0 \geq 1 - y w^\top x_i$

  - Define $C_i \triangleq \{\boldsymbol{w} \in \mathbb{R}^d \mid y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1\}$.
  - For any $i \in [m]$, $t \in [0,1]$, and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in C_i$:

    $\cancel{tz}\quad \alpha u + (1-\alpha) v$

    - We start by showing C_i is convex, that is, $t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2 \in C_i$
    - This happens iff $y_i(t\boldsymbol{w}_1^\top + (1-t)\boldsymbol{w}_2^\top)\boldsymbol{x}_i \geq 1$

Lemma 1:
a linear function is convex

$$y_i(t\boldsymbol{w}_1^\top + (1-t)\boldsymbol{w}_2^\top)\boldsymbol{x}_i = t\underbrace{y_i\boldsymbol{w}_1^\top\boldsymbol{x}_i}_{\geq 1} + (1-t)\underbrace{y_i\boldsymbol{w}_2^\top\boldsymbol{x}_i}_{\geq 1} \geq t + (1-t) = 1$$

  - The set $C = \bigcap_{i=1}^{m} C_i$ is the intersection of convex set $\Longrightarrow$ C is convex.

Back to the Hard-SVM formulation

- We showed that:
  - $f(\boldsymbol{w})$ is convex
  - C is convex

$$\underset{\boldsymbol{w} \in C}{\operatorname{argmin}} \; f(\boldsymbol{w})$$

- Convince yourself: if we restrict a convex function to a convex subset, then it is a convex function, i.e., $f|_C$ is convex.
  - Hint: if $(tx_1 + (1-t)x_2) \in C$ then $f|_C(tx_1 + (1-t)x_2) = f(tx_1 + (1-t)x_2)$

- Corollary: Hard-SVM is convex.

Soft-SVM is also convex

- Soft-SVM is also convex!
- Use the hinge-loss formulation

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \lambda ||\boldsymbol{w}||_2^2 + \frac{1}{m}\sum\nolimits_{i\in[m]} \ell_{hinge}(\boldsymbol{w}, \boldsymbol{x}_i)$$

- In Short HW 3: prove that the Soft-SVM objective is convex.

GRADIENT DESCENT

An iterative algorithm for convex optimization

Gradient Descent (GD)

- An iterative minimization method.
- Asks "what is the steepest way down?" and steps in that direction.

Pseudo code

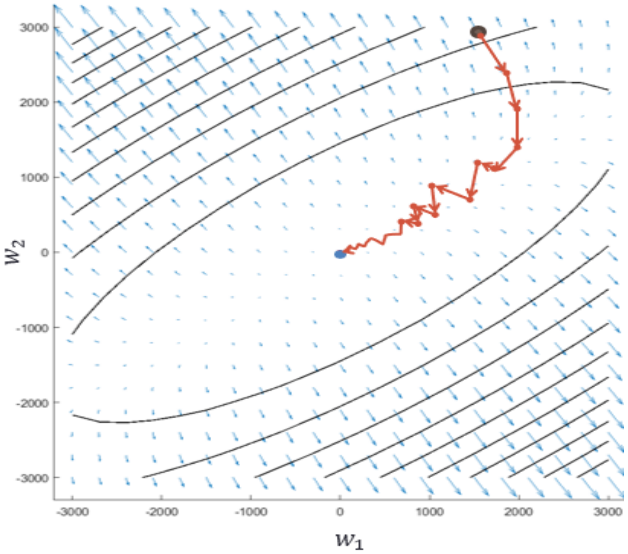
Choose learning rate η

Initialize a random starting point x_0

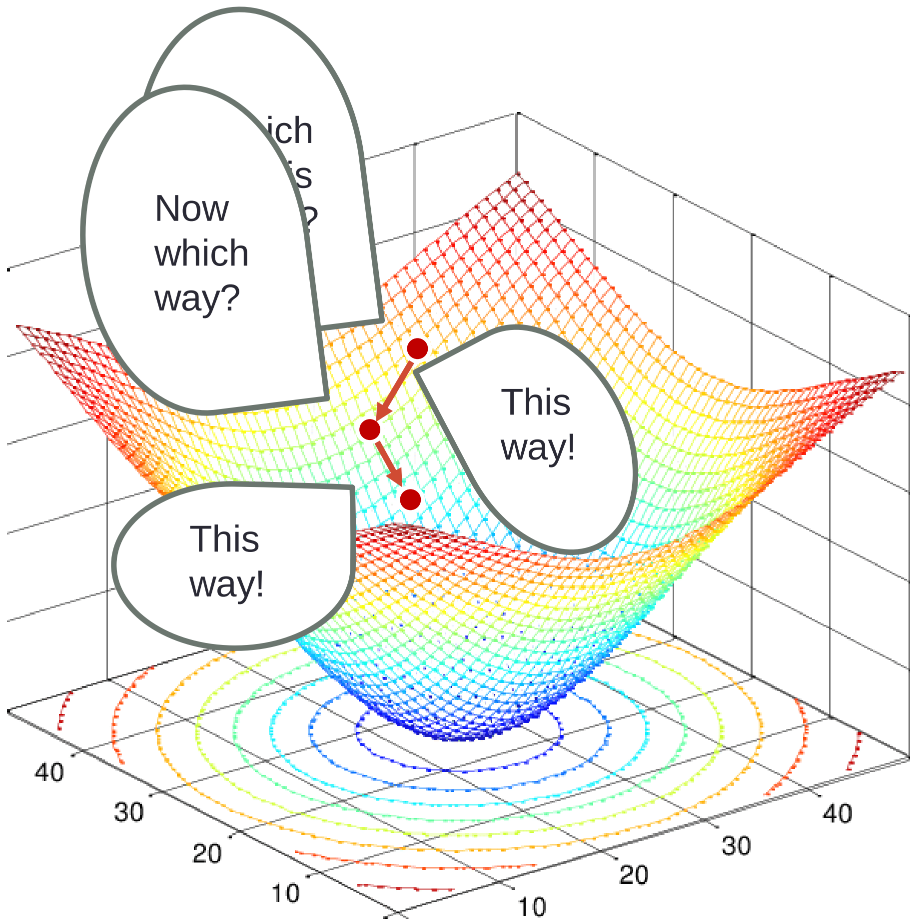
For i=1, …, num_iters:

Calculate $\nabla f(x_i)$

Set $x_{i+1} = x_i - \eta \cdot \nabla f(x_i)$



Gradient Descent (GD)



Source: [Barbarosou and Maratos 2011]

Pseudo code

Choose learning rate η

Initialize a random starting point x_0

For i=1, …, num_iters:

  Calculate $\nabla f(x_i)$

  Set $x_{i+1} = x_i - \eta \cdot \nabla f(x_i)$

Gradient Descent (GD)

- An iterative minimization method.
- Asks "what is the steepest way down?" and steps in that direction.

- Guaranteed to converge to a local minimum when the learning rate is small enough (more on that later).
- Remember that for a convex function, any local minimum is global!

Pseudo code

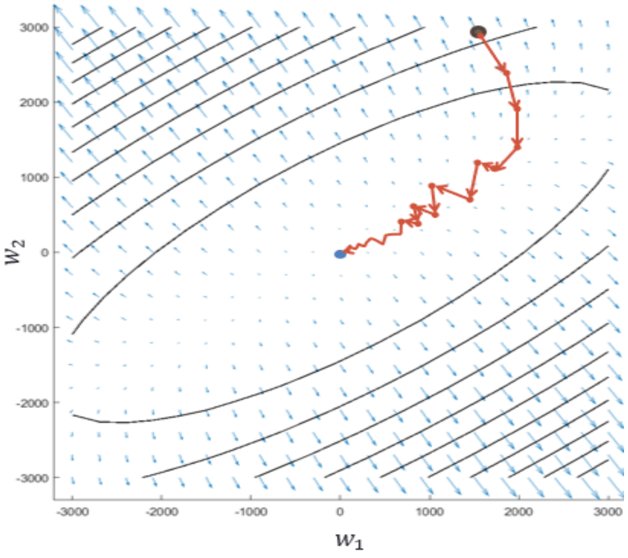
Choose learning rate η

Initialize a random starting point x_0

For i=1, …, num_iters:

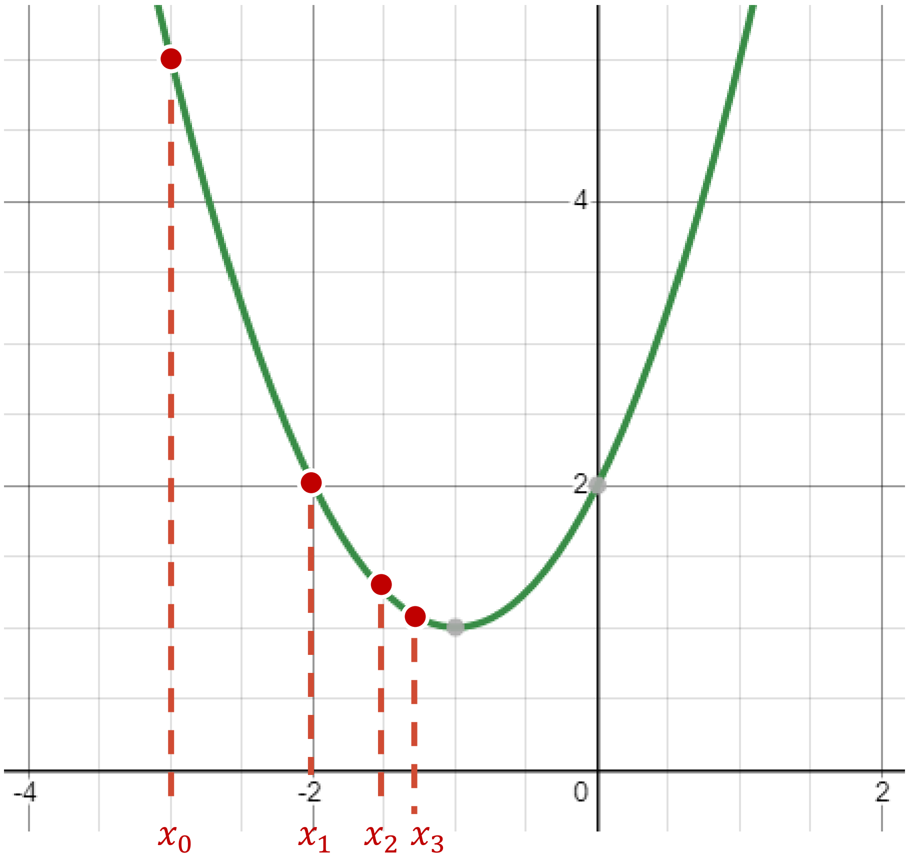
  Calculate $\nabla f(x_i)$

  Set $x_{i+1} = x_i - \eta \cdot \nabla f(x_i)$



Gradient descent in 1D

$f(x) = x^2 + 2x + 2$ $\qquad$ $\nabla f(x) = 2x + 2$



Pseudo code

Choose LR η and starting point x_0

For i=1, …, num_iters:

Calculate $\nabla f(x_i)$

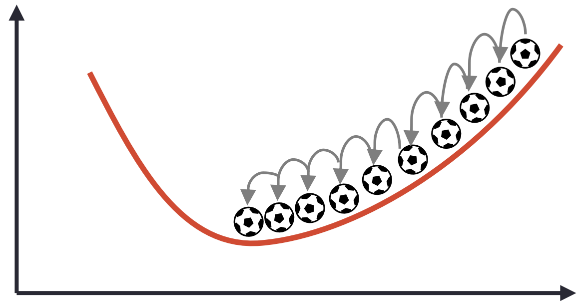
Set $x_{i+1} = x_i - \eta \cdot \nabla f(x_i)$

$\boldsymbol{init}$: $\eta = 1/4$,	$x_0 = -3$	
i	x_i	$\nabla f(x_i)$
0	-3	-4
1	-2	-2
2	-1.5	-1
3	-1.25	
⋮		

The learning rate (step size)

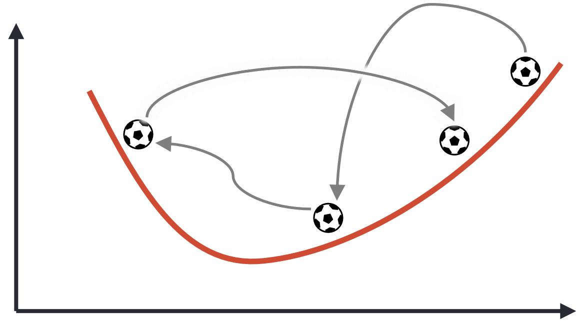
- The learning rate η controls the rate of convergence to the minimum.
- GD is like pushing a ball down a valley, and η is the push force.

Too small



Long time to converge

Too large



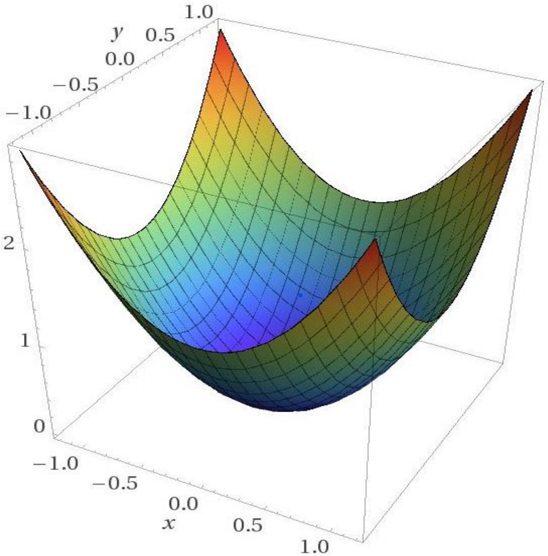
Miss the minimum,
might never converge

- Let's play around with this parameter: Google Colab

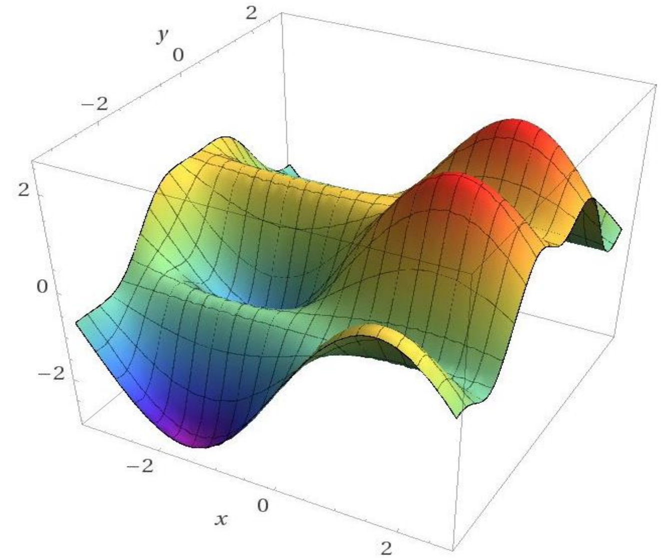
GD on non-convex functions

- Can we perform GD on non-convex functions?
  - Yes! But can only hope to converge to a local minimum.

Convex landscape
(e.g., SVM)



Non-convex landscape
(e.g., Deep learning)



Source: VProexpert

Summary

- Defined convexity of sets and functions.
- Saw several tools to prove convexity.
- SVM is a convex optimization problem.
- We can converge to a global minimum of any convex function using gradient descent with a small enough learning rate.