Lecture 8. Function convexity and Gradient descent

COMP 551 Applied machine learning

Yue Li
Assistant Professor
School of Computer Science
McGill University

September 27, 2022

Outline

Learning objectives

Understanding the basic ideas of

- Function convexity
- Gradient descent as general algorithm
- Stochastic gradient descent method of momentum
- Adaptive learning rate

Numerical optimization is the workhorse of machine learning algorithms

In this lecture we consider:

- Continuous variables
- Unconstrained
- Convexity testing
- Local optima
- Analytic gradient
- Stochastic
- Smooth error surface

In this lecture we do **not** consider:

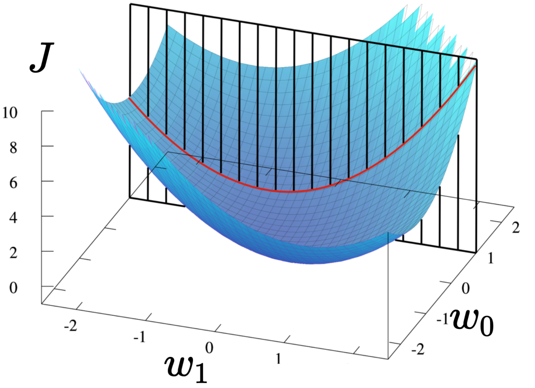
- Discrete variables
- Constrained (next lecture)
- Global optima
- Proximal gradient
- Analytic Hessian
- Non-smooth error surface

Gradients

For a cost function involving multiple variables, gradient is the partial derivative of the function w.r.t. each variable while fixing other variable.

For example, in a two-variable regression, the gradient of the coefficient w_1 is:

$$\frac{\partial}{\partial w_1} J(w_0, w_1) = \lim_{\epsilon \to 0} \frac{J(w_0, w_1 + \epsilon) - J(w_0, w_1 - \epsilon)}{2\epsilon}$$



Gradients

In a more generally of D variables setting, we have

$$\frac{\partial}{\partial w_d} J(w_0, w_1, \ldots, w_{D-1}) = \lim_{\epsilon \to 0} \frac{J(w_d + \epsilon, \mathbf{w}_{\setminus d}) - J(w_d - \epsilon, \mathbf{w}_{\setminus d})}{2\epsilon}$$

where $\mathbf{w}_{\setminus d}$ denotes all coefficients except for coefficient w_d.
Gradients in matrix notation:

$$\nabla J(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_0} J(\mathbf{w}) \\ \frac{\partial}{\partial w_1} J(\mathbf{w}) \\ \vdots \\ \frac{\partial}{\partial w_{D-1}} J(\mathbf{w}) \end{bmatrix}$$

General gradient descent algorithm

Gradient descent is an iterative algorithm for optimization.

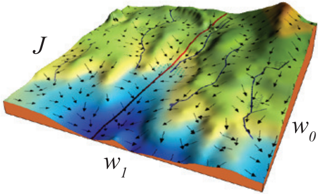
Algorithm 1 GradientDecent($\alpha = 0.005$, Convergence Criteria)

1: Initialize coefficients $\mathbf{w}^{(0)}$
2: **while** convergence criterion is not met **do**
3: $\quad \mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \alpha \nabla J(\mathbf{w})$ // *update using gradient*
4: **end while**

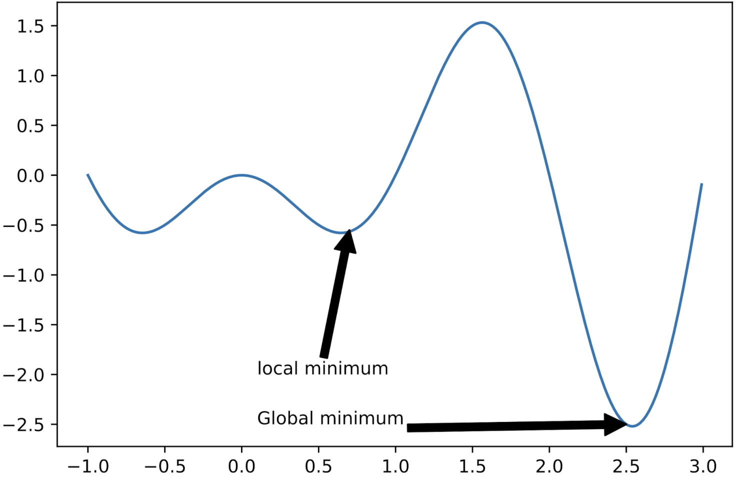
where

- $\mathbf{w}^{(t)}$: weights at the t^{th} iteration
- α: learning rate
- $\nabla J(\mathbf{w}) = \left[\frac{\partial}{\partial w_0} J(\mathbf{w}), \frac{\partial}{\partial w_1} J(\mathbf{w}), \ldots \frac{\partial}{\partial w_{D-1}} J(\mathbf{w})\right]^\top$ are the gradients

Steepest decent into a *global or local optimum* of the error surface



How to determine if we can always find *the global optimum* not local optimum given a loss function?

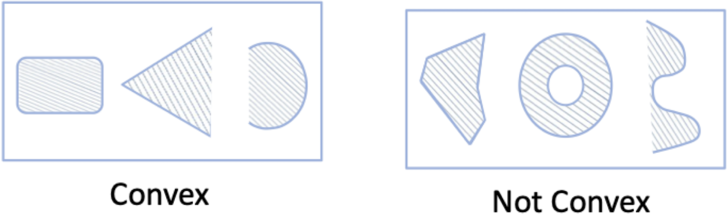


Convex sets

$\mathcal{S}$ is a **convex sets** if, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{S}$, we have

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{x}' \in \mathcal{S}, \forall \lambda \in [0,1]$$

In English: if we draw a line from $\mathbf{x}$ to $\mathbf{x}'$, all points on the line lie inside the set.

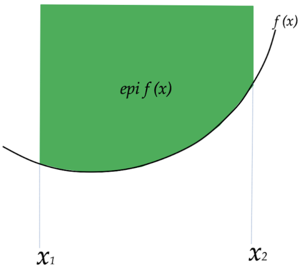


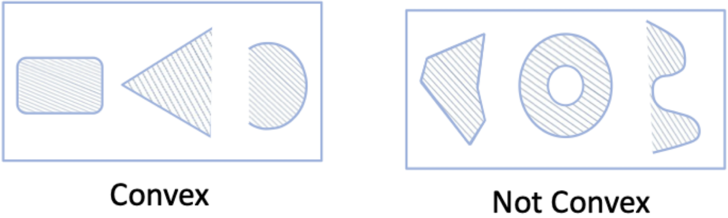
Convex functions

- *epigraph* is defined by a set of points above the function
- A function $f(\mathbf{x})$ is *convex* if its epigraph is a convex set
- Equivalently, a function $f(\mathbf{x})$ is convex if

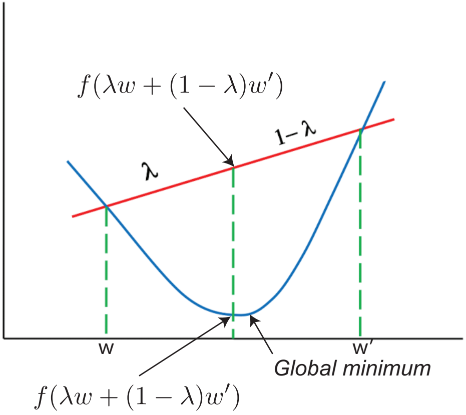
$$f(\lambda w + (1-\lambda)w') \leq \lambda f(w) + (1-\lambda) f(w')$$

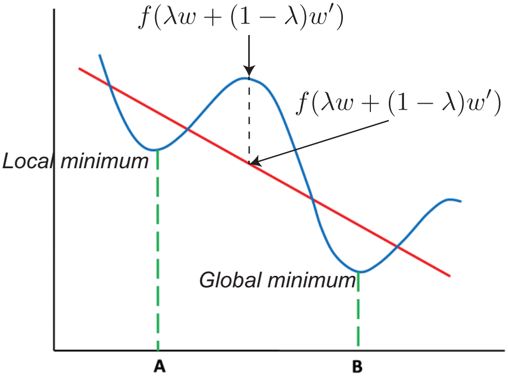
where $f(\lambda w + (1-\lambda)w')$ are points on the curve and $\lambda f(w) + (1-\lambda) f(w')$ are points on the red line



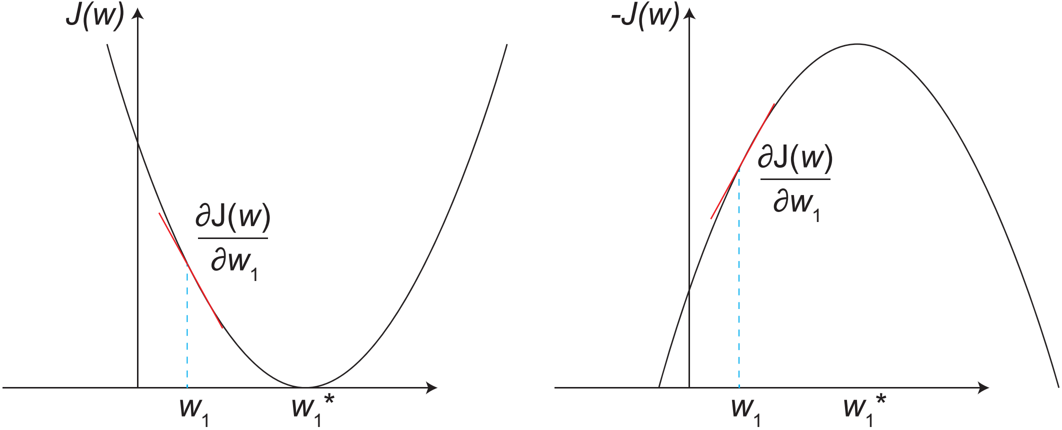


Examples of a convex and a non-convex function





Negative of a convex function is a concave function



For example, minimizing sum of squared error is a convex function and its negative is a concave function and is equivalent to the log likelihood of a Gaussian (Lec 5).

Why do we care about convexity?

- Gradient descent (GD) can find the global minimum on a convex function.
- GD on a non-convex function may find local minimum (i.e., ultimately resulting in low prediction accuracy)

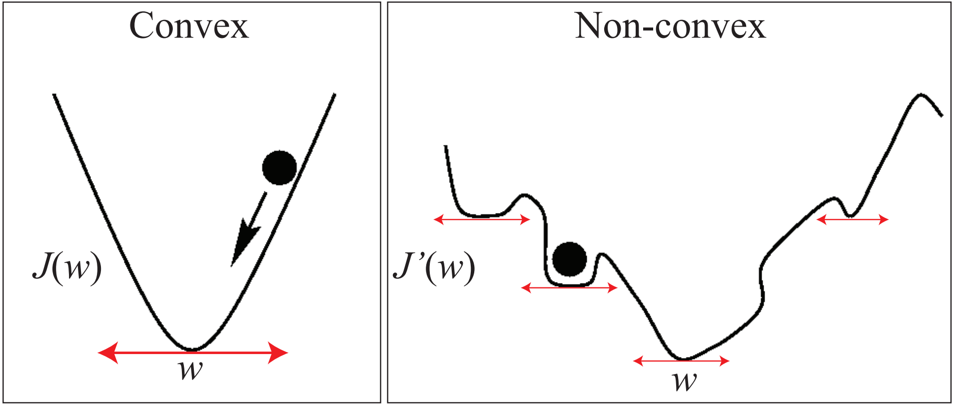


Image source source

Recognizing convex functions

The principled way to test whether a function $f(x)$ is convex is by taking the second order derivative of the function:

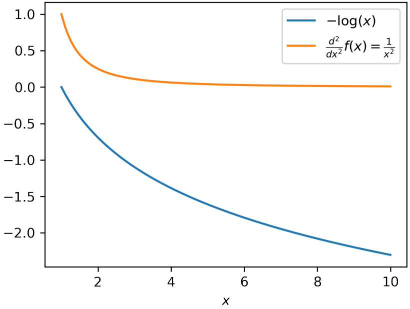
$$\frac{\partial}{\partial x}\frac{\partial f(x)}{\partial x} = \frac{\partial^2 f(x)}{\partial x^2} \equiv \nabla^2 f(x)$$

For example, the first-order derivative of the negative natural log function $f(x) = -\log(x)$ is:

$$\frac{\partial f(x)}{dx} = -\frac{1}{x}$$

and its second-order derivative is:

$$\frac{d^2 f(x)}{dx^2} = \frac{1}{x^2}$$



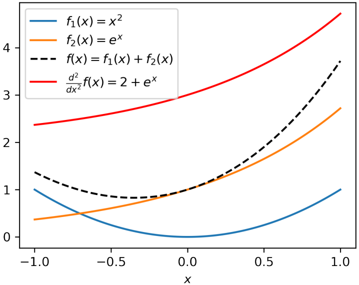
which is *always positive* for $\forall x \in \mathbb{R}$.

Some more examples of 1d convex functions $f(x)$: $-\sqrt{x}$, x^2, e^x, $x\log(x)$

Sum and max of convex functions are still convex functions

Sum of convex functions is convex.

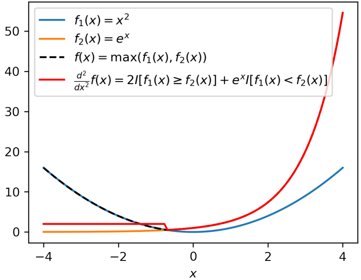
- Example 1. $f_1(x) = x^2$ and $f_2(x) = e^x$ are both convex and $f(x) = f_1(x) + f_2(x)$ is also convex.



- Example 2. sum of squared errors is a convex function: $J(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - y||_2^2 = \sum_n (\mathbf{x}^{(n)}\mathbf{w} - y^{(n)})^2$

Maximum of convex functions is convex.

- Example 1. $f_1(x) = x^2$ and $f_2(x) = e^x$ are convex, $f(x) = \max(f_1(x), f_2(x))$ is also convex.



- Example 2. $f_1(y) = y^4$ is convex and maximum of $f_1(y)$ is also convex:

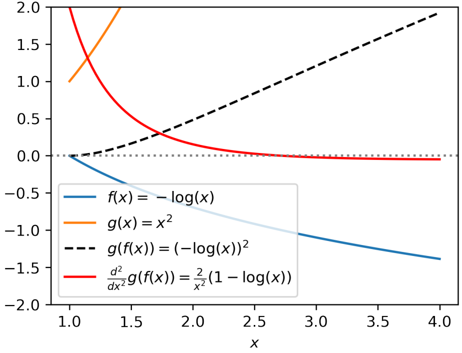
$$f(y) = \max_{x \in [0,2]} x^3 y^4 = 8y^4$$

Composition of convex functions may or may not be convex

For example, $f(x) = -\log(x)$ and $g(x) = x^2$ are convex but not:

$$g(f(x)) = (-\log(x))^2$$

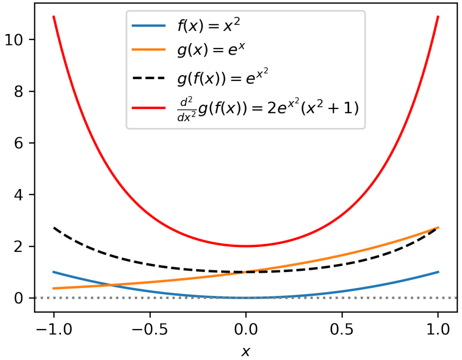
$$\nabla^2 g(f(x)) = \frac{2}{x^2}(1 - \log(x)) < 0 \text{ if } x > e$$



If $f(x)$, $g(x)$ are convex, and $g(x)$ is **non-decreasing**, then $g(f(x))$ is convex. e.g.,

$$f(x) = x^2, g(x) = e^x, g(f(x)) = e^{x^2}$$

$$\nabla^2 g(f(x)) = 2e^{x^2}(x^2 + 1) > 0 \, \forall x \in \mathbb{R}$$



Verify if cross-entropy is a convex functions

Coming soon . . .

Summary

- a