Recitation 1
Elements of the Good Life: Calculus and Convexity

10-315: Introduction to Machine Learning

Fall 2020

1 Calculus

We all are comfortable with single-variable calculus(I hope).

Most often we will be dealing with multi-variable scalar valued functions $f : \mathbb{R}^n \to \mathbb{R}$. Their derivative Df is defined as the gradient:

Definition 1.1. Gradient ∇ If $f : \mathbb{R}^n \to \mathbb{R}$ (is sufficiently nice) then we have the gradient as the derivative

$$\nabla f = \langle \frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n} \rangle$$

ie. its just a list of single variable derivatives in the direction of the axes. Note this is a vector when evaluated at a point.

Ex: Square l^2 norm

Definition 1.2. Square l^2 norm For $x \in \mathbb{R}^n$ define the square l^2 norm of x, denoted $||x||_2^2$ as

$$||x||_2^2 = \sum x_i^2$$

1. What is the gradient of the $||x||_2^2$ at arbitrary $x \in \mathbb{R}^n$?

   Compute the ith partial $\frac{\partial f}{\partial x_i}(x) = 2x_i$ so the gradient is

   $$\nabla f(x) = \langle 2x_1, .., 2x_n \rangle$$

Definition 1.3. Hessian H For (sufficiently nice) $f : \mathbb{R}^n \to \mathbb{R}$ define the hessian of f($D^2 f$) as

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

The ith column can be thought of as the gradient of the ith coordinate of the gradient of f. Note this is a matrix when evaluated at a point, and in any situation we will encounter the partial will commute, ie. $\partial_1 \partial_2 f = \partial_2 \partial_1 f$ so it is symmetric.

Ex:

1. What is the Hessian of the $||x||_2^2$? at arbitary $x \in \mathbb{R}^n$

   We already computed the gradient. So to compute the hessian we may just compute each column, ie. the gradient of each component of ∇f.

   Recall $(\nabla f(x))_i = 2x_i$. This has gradient

   $$\langle 0, ..., 2, ..., 0 \rangle$$

   where the ith coordinate is nonzero. This implies the hessian is

   $$\begin{bmatrix} 2 & 0 & 0 & \dots \\ 0 & 2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 2 \end{bmatrix}$$

   Alternatively we could have just computed $\partial_i \partial_j f$ which is 2 for very i,j.

Moral: If you can do single-variable calculus you can do multivariable calculus.(At least in this class).

2 Convexity

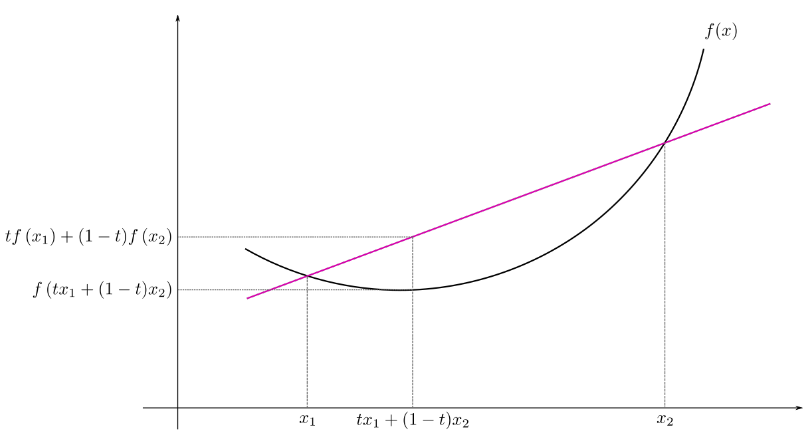
Now we address an extremely nice class of functions: those which are *convex.*

Definition 2.1. Convex Fuctions $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\forall t \in [0, 1]$, $x, y \in \mathbb{R}^n$,

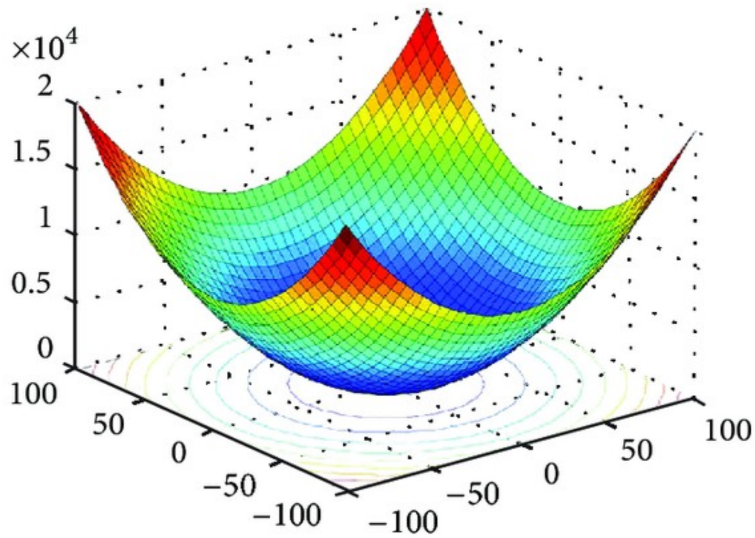
$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

Also (erroneously) known as "concave-up".

Geometry of convex functions in one dimension:



In n-dimensions:



In a very strong sense sub-linear, ie. overapproximated by its gradient.

An equivalent definition(for sufficiently differentiable functions) requires the hessian of f to be *positive semi-definite*

Definition 2.2. Convex Fuctions $f : \mathbb{R}^n \to \mathbb{R}$ is convex for all $x \in \mathbb{R}^n$ the hessian $Hf(x)$ is positive semi-definite.

Definition 2.3. Positive Semi-Definite Matrix A matrix $H \in \mathbb{R}^{n\times n}$ is positive semi definite if for all $x \in \mathbb{R}^n$, $x^T Hx \geq 0$

(The bilinear form induced by H satisfies positivity). We mentioned in our discussion of hessians that our hessians will almost always be symmetric. In the case of a symmetric matrix we have the following equivalence

Theorem 1. Symmetric Positive Semi-Definite Matrices If $H \in \mathbb{R}^{n\times n}$ symmetric then it is positive semi-definite $\iff$ all its eigenvalues are ≥ 0.

Proof. Not super relevant but a good exercise/refresher in linear algebra. If you get stuck can find on math stack exchange. □

This condition is often easier to check and thus good to know.

Further note that in one dimension showing convexity amounts to showing $f''(x) \geq 0$.

2.1 Determining if a Function is Convex

Ex:

1. Show $f(x) = ||x||_2^2$ is convex

   Recall the hessian of f is

$$\begin{bmatrix} 2 & 0 & 0 & \dots \\ 0 & 2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 2 \end{bmatrix}$$

Note it has only one eigenvalue, 2. And $2 \geq 0$ so we know f convex.

Alternatively for $x \in \mathbb{R}^n$, $x^T H x = x^T 2x = 2x^T x \geq 0$ since $x^T x \geq 0$. This shows the definition directly

There are many other methods of showing convexity:

Theorem 2. If f, g are convex then $f + g$ convex.

Proof. Linearity of the derivative. □

Theorem 3. If f convex then αf convex for $\alpha \in \mathbb{R}^+$.

Proof. Linearity of the derivative. □

In the one-dimensional case $f, g : \mathbb{R} \to \mathbb{R}$:

Theorem 4. If f and g are convex functions and g is non-decreasing, then $g(f(x))$ is convex.

Proof. Chain rule □

Theorem 5. If f is concave and g is convex and non-increasing then $g(f(x))$ is convex

Proof. Chain rule □

2.2 Nice Properties of Convexity

Often in machine learning we are seeking to minimize some objective function/error function $f : \mathbb{R}^n \to \mathbb{R}$ in order to get a best fit for our model. In general this is very hard even if f is sufficiently differentiable(there could be many local minima/maxima, but we want the global).

However if our objective f is convex then we can find the extrema easily!

Ex:

1. Classify the extrema of $f(x) = ||x||_2^2$ on $\mathbb{R}^n$

   Recall that if x_0 is a minimum or maximum then $\nabla f(x_0) = 0$. So if $\nabla f(x_0) = \langle 2x_1, ..., 2x_n \rangle = 0$ it must be $x_0 = 0$ is a unqiue extrema!(In this unconstrained problem).

2. What does the hessian H of f tell us about the extrema at $x_0 = 0$

   $f(x_0) = f(0) = 0$ must be a global minimum, since $Hf(x_0) > 0$. This is obvious but I'm trying to demonstrate a more general principle.

Theorem 6. Suppose $x_0 \in \mathbb{R}^n$ is s.t. $\nabla f(x_0) = 0$. Then x_0 is a local minimum for f if $Hf(x_0) > 0$. Further it is a local maximum if $Hf(x_0) < 0$.

In the single variable case this goes by the second derivative test:

Theorem 7. Suppose $x_0 \in \mathbb{R}^n$ is s.t. $f'(x_0) = 0$. Then x_0 is a local minimum for f if $f''(x_0) > 0$. Further it is a local maximum if $f''(x_0) < 0$.

So this immediately tells us that all convex functions must have only minima. **Furthermore all convex funcions have a unique global minimum** if one exists. We can show this clearly in the single variable case:

1. Let $f : \mathbb{R} \to \mathbb{R}$ be convex and twice differentiable. Argue if it has a minima, it is unique.

   Recall for any extrema z, $f'(z) = 0$. Further since $f''(x) \geq 0$ for all x we know f' is nondecreasing and hence by IVT has at most one $f'(z) = 0$ occurs at most once. Further it must be a minima as f is convex and $f''(z) \geq 0$.

To find the global minimizer of a convex function f set $\nabla f(x) = 0$ and solve.