Calculus review

Mithun Nallana

IIIT Hyderabad

mithun.babu@research.iiit.ac.in

May 11, 2019

Overview

Why?

Goals:

- Review important material before diving-in.
- Collect all notations at one place.
- Able to read books and seminal papers and actively participate in projects.

Spend some time after class to connect the dots.

"Before a man studies Zen, to him mountains are mountains and waters are waters; after he gets an insight into the truth of Zen through the instruction of a good master, mountains to him are not mountains and waters are not waters; but after this when he really attains to the abode of rest, mountains are once more mountains and waters are waters." - D.T.Suzuki

Gradient

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

We can now extend this definition to multi-variate functions $f : \mathrm{I\!R}^{\mathrm{n}} \to \mathrm{I\!R}$,

$$\frac{\partial f(x_1, x_2, \dots x_n)}{\partial x_1} = \lim_{h_1 \to 0} \frac{f(x_1 + h_1, x_2, x_3, \dots x_n) - f(x_1, x_2, \dots x_n)}{h_1}$$

Similarly we have,

$$\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}$$

Gradient

We stack them together in a vector,

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Hessian

Hessian of a $f : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\nabla^2 f = H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Notice that this is a symmetric matrix.

Examples

$$f(x, y, z) = 3x^2yz$$

Find gradient and hessian.

Model and Error

We have a model $f(P)$ to estimate a measurement X

$$X = f(P)$$

Here P are the parameters for model and X can be an property of environment that can be measured through sensors directly.

Our goal is to find a $\hat{P}$ such that,

$$||\epsilon||_2 = ||f(\hat{P}) - X||_2$$

is minimized.

Notice that $f(\hat{p})$ can be either linear or non-linear.

Model and Error

Few pressure sensors, inertial sensors can be represented as linear models in working range.

$$f(P) = AP$$

We have measured some reading from the sensor, b.

Now there are many possible cases. In few cases finding the parameters will be easy and few others it is not.

$$AP = b$$

$A \in \mathrm{I\!R}^{\mathrm{m \times n}}$ and $P \in \mathrm{I\!R}^{\mathrm{n}}$

Exact solution:

- If $m < n$, we have many solutions. They form a vector space.
- If $m = n$, we have either a unique solution or no solution.
- If $m > n$, we have either a unique solution or no solution.

Linear equations

No solution case:

- This case occurs when b doesn't lie in the column space of A.
- One way is to find a nearest vector in column space of A that is close to b.
- We have to find $AP - b$ which is orthogonal to column space of A.

$$A^T(AP - b) = 0$$
$$A^T AP = A^T b$$

- This system will have a solution as both right and left side are in column space of A^T.
- These are normal equations and $(A^T A)^{-1} A^T$ is called pseudo inverse.

Non-Linear

Sometimes we can have a highly non-linear sensor model.

However, we have some estimate of what $\hat{P}_0$ through another inaccurate sensor.

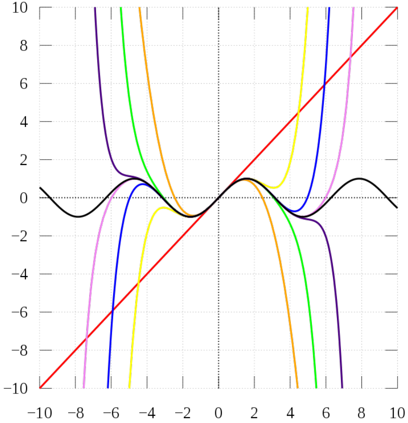
Our goal is to find the best value of $\hat{P}$ that fits the measurement data well.

So, there is a need of iterative optimization techniques.

Taylor series

For a ∞ differentiable real valued function, Taylor series can be written as,

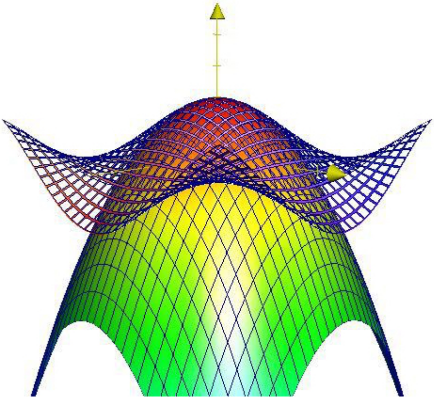
$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \dots$$



Taylor series

This has been generalized to multivariate functions,

$$f(x) = f(a) + \frac{\nabla f(a)^T (x-a)}{1!} + \frac{(x-a)^T \nabla^2 f(a)(x-a)}{2!} + \dots$$



Example

Taylor series expansion of $cos(x)$ at $x = 0$

$$cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$$

Let's estimate,

$$\begin{aligned} cos(0) &= 1 \\ cos(0.2) &= 0.9800665 \\ cos(0.2) &\approx 1 \\ cos(0.2) &\approx 0.98 \\ cos(0.2) &\approx 0.9800666 \\ cos(0.2) &\approx 0.9800665 \end{aligned}$$

Cost function

Our goal is to minimize,

$$||\epsilon||_2 = ||f(P) - X||_2$$

Let our cost/error function be $g(P)$

We have an initial estimate of $P = P_0$

Using taylor series (2^{nd} order), we can approximate cost function as,

$$g(P_0 + \delta P) = g(P_0) + \nabla g^T(\delta P) + \delta P^T(\nabla^2 g)\delta P$$

First-order method

If we look until 1^{st} order approximation,

$$g(P_0 + \delta P) = g(P_0) + \nabla g^T(\delta P)$$

The best possible choice of δP is $-\lambda \nabla g$

$$g(P_0) - \lambda * \nabla g^T \nabla g$$

This method is called as gradient descent.

Gradient descent

In case of above $g(P)$ We have,

$$g(P_0) = f(P_0) - X$$

Assuming that sensor behaves linearly at P_0

$$f(P_1) = f(P_0 + \delta P) = f(P_0) + J^T \delta P$$

Where $J = \frac{\partial f}{\partial P}$

Our goal is to minimize,

$$\begin{aligned} g(P_1) &= f(P_1) - X \\ &= f(P_0) + J^T \delta P - X \\ &= g(P_0) + J^T \delta P \end{aligned}$$

Gradient descent

$$||g(p_1)||_2$$

$$J^T \delta P = -g(P_0)$$
$$JJ^T \delta P = -J(g(P_0))$$

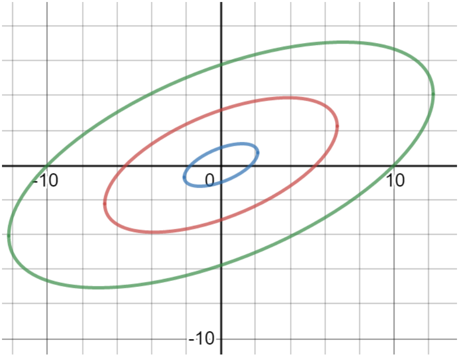
This method is known as Gauss-Newton method.

Gradient descent

Observation,

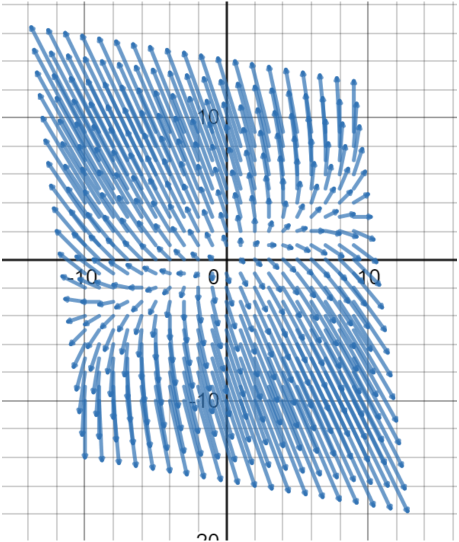
$$f(x,y) = x^2 + 3y^2 - 2xy$$
$$\nabla f = \begin{bmatrix} 2x - 2y \\ 6y - 2x \end{bmatrix}$$

Contour lines:



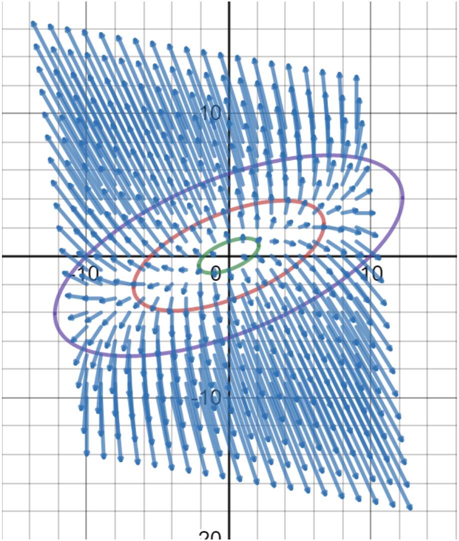
Gradient descent

Vector field:



Gradient descent

Vector field:



Gradient descent

Find a direction that has maximum increase in function.

Directional Derivative:

$$\nabla_{\vec{v}} f = \lim_{h \leftarrow 0} \frac{f(x + h\vec{v}) - f(x)}{h}$$

We can divide increment of function into components:

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}, ||v||_2 = 1$$

Gradient descent

Moving along v_1, We have

$$p_1 = h * v_1 * (\frac{\partial f(x_1, x_2, \dots x_n)}{\partial x_1})$$

Similarly,

$$\begin{aligned} p_2 &= h * v_2 * (\frac{\partial f(x_1 + hv_1, x_2, \dots x_n)}{\partial x_2}) \\ p_3 &= h * v_3 * (\frac{\partial f(x_1 + hv_1, x_2 + hv_2, \dots x_n)}{\partial x_3}) \\ &\vdots \end{aligned}$$

Final value of function is,

$$f(x_1, x_2 + \dots x_n) + p_1 + p_2 + p_3 + \dots + p_n$$

Gradient descent

Look at $p_1, p_2, \dots p_n$ closely,

$$p_1 = h * v_1 * \left(\frac{\partial f(x_1, x_2, \dots x_n)}{\partial x_1}\right)$$

$$p_2 = h * v_2 * \left(\frac{\partial f(x_1, x_2, \dots x_n)}{\partial x_2}\right)$$

$$\vdots$$

so, $p_1 + p_2 + \dots p_n$ is

$$h * \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Gradient descent

$$p_1 + p_2 + \dots p_n = \nabla f^T \vec{v}$$

Our goal is to maximize the sum which is possible when $\vec{v}$ is along ∇f.

Thus our defined gradient direction is the direction of maximum ascent.

LevenbergMarquardt

Gradient descent,

$$\delta P = -\lambda \nabla g$$

Gauss-Newton,

$$(JJ^T)\delta P = -J(g(P_i))$$

LevenbergMarquardt,

$$((JJ^T) + \lambda I)\delta P = -J(g(P_i))$$

Gradient descent

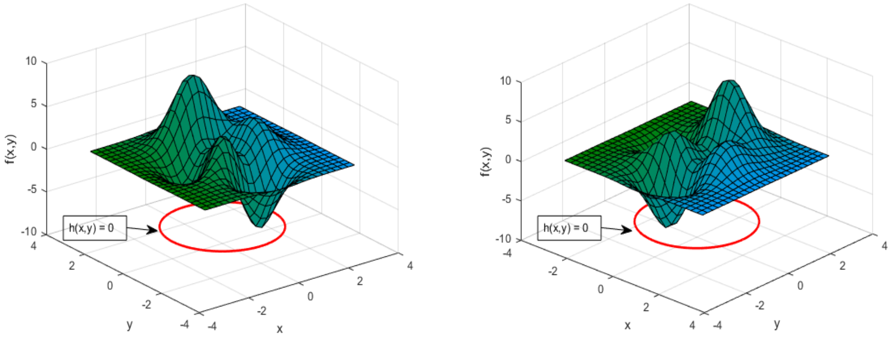
Given

$$\underset{x}{\text{minimize}} \quad f(x)$$
$$x \in \mathbb{R}^{\mathrm{n}}$$

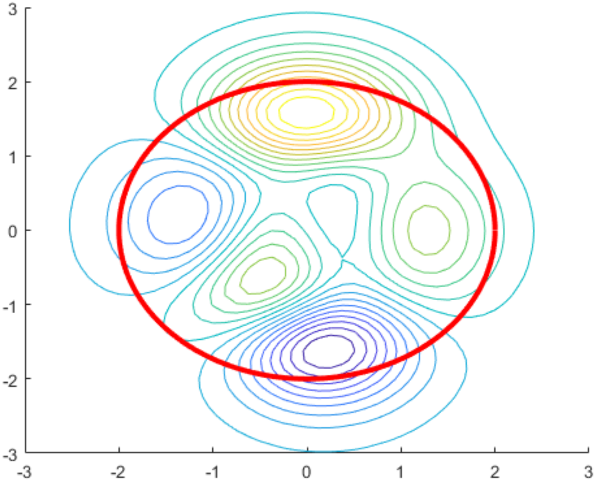
Constrained problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) = 0, \; i = 1, \ldots, m. \\ & x \in \mathbb{R}^{\mathrm{n}} \end{aligned}$$

Constrained problem



Constrained problem



Constrained problem

$$\nabla f(x, y) = \nu \nabla h(x, y)$$
$$h(x, y) = 0$$

Written in other way:

$$\underset{x,y,\nu}{\text{minimize}} \quad f(x, y) + \nu h(x, y)$$

We have converted a constrained optimization problem to an unconstrained one.

Constrained problem

$$\underset{x}{\text{minimize}} \quad f(x) + \sum_{i=1}^{m} I_0(h_i(x))$$
$$x \in \mathbb{R}^{\mathrm{n}}$$

Here, I_0 is the indicator function defined as,

$$I(u) = \begin{cases} 0 & u = 0 \\ \infty & \text{otherwise} \end{cases}$$

I_0 function can be understood as high displeasure to constraint violation. As constraint gets far away from zero.

Lagrange multiplier can be seen as a smooth approximation to I_0.

References

- MITOCW 18.02 course
- Muti view geometry by Hartley and Zisserman.
- Convex optimization by Stephen Boyd.
- Notes by Hal Daume
- Khan Academy Notes and Videos
- Lagrange Multipliers notes by Yan Bin jia
- Wikipedia

The End