Fundamentals of Unconstrained Optimisation

Saurav Samantaray

Department of Mathematics

Indian Institute of Technology Madras

January 29, 2024

Unconstrained Optimisation

- Minimise an objective function that depends on real variables.
- No restriction on the values of these variables (no constraints).

Mathematical Formulation:

$$\begin{gathered} \min_x f(x) \\ \text{where, } x \in \mathbb{R}^n, n \geq 1. \\ f : \mathbb{R}^n \to \mathbb{R} \quad \text{is smooth} \end{gathered} \tag{1}$$

In a real world scenario

- The objective function "f" might not be known globally everywhere.
- Ideally, may have finitely many values of "f" or some derivatives of "f".
- Any information for "f" at arbitrary points usually do-not come very cheaply.
- Therefore, one should prefer for algorithms which do-not demand the same, unnecessarily.

Example

- Suppose we are trying to find a curve that fits some experimental data.
- (t_i, y_i), y_i signal is measured at time t_i.
- Let's assume based on the knowledge of the phenomenon under study we have the understanding that the signal has exponential and oscillatory behaviour of certain types.

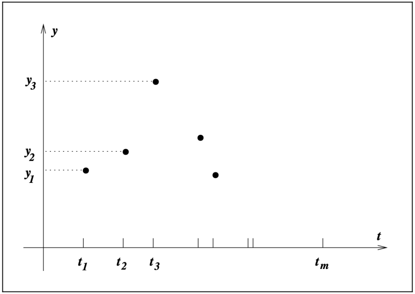

Figure: Least squares data fitting problem.

Example

- Choose the model function as

$$\phi(t,x) = x_1 + x_2 e^{-(x_3-t)^2/x_4} + x_5 \cos(x_6 t)$$

  where x_i's are the parameters of the model.
- What we want is the model should fit the observed data y_j, as closely as possible.
- Let $x = (x_1, x_2, x_3, x_4, x_5, x_6)$,
  We define the residual for each y_j as

$$r_j = y_j - \phi(t_j, x), \qquad j = 1, \ldots, m.$$

- We define the objective function as

$$\min_{x \in \mathbb{R}^6} f(x) = r_1^2(x) + \ldots + r_m^2(x)$$

This is a non-linear least square problem, a special case of unconstrained optimisation.

Example

- Note that the equation of the objective function appears quite expensive even for small number of variables

$$n = 6$$

- say, if the no. of measurements i.e. $m = 10^5$, then the evaluation of f becomes quite a computational expense.

Lets Gain Some Perspective!!

- Suppose for a given set of data the optimal solution to the previous problem is approximately

$$x^* = (1.1, 0.01, 1.2, 1.5, 2.0, 1.5)$$

  and the corresponding function value is $f(x^*) = 0.34$.
- As at the optimal point the objective is non-zero there must be some discrepancy between the function values and the observations made.

Some Perspective

- i.e. y_j and $\phi(t_j, x^*)$ aren't the same for some or many $(y_j, t_j) \longleftrightarrow \phi(t_j, x^*)$
- The model hasn't produced all the data points correctly as

$$f(x^*) \neq 0$$

- Then how to know x^* is indeed a minimiser of f?
- In the sense that how to know which all points one should go close to or not?
- To answer this question, we need to define the term "solution and explain how to recognise solutions.

What is a solution?

A point x^* is a global minimiser of f if

$$f(x^*) \leq f(x) \qquad \forall\, x \in \mathbb{R}^n$$

or in the domain of interest.

- It would be the most ideal scenario if we could find a global minimiser.
- It might be difficult to get a global minimiser, owing to the limited (or local) knowledge of f.
- Most algorithms are only able to find a local minimiser.

What is a solution?

Local Minimiser

A point x^* is called a local minimiser, if there is a neighbourhood $\mathcal{N}$ of x^* such that

$$f(x^*) \leq f(x) \quad \forall\, x \in \mathcal{N}$$

- It's a points that achieves the smallest value of f in its neighbourhood.

Weak Local Minimiser

$$f(x^*) \leq f(x) \qquad x \in \mathcal{N}$$

Strict (Strong) Local Minimiser

$$f(x^*) < f(x) \qquad x \in \mathcal{N},\ x \neq x^*$$

Example

- For a constant function $f(x) = 2$ every point is a weak local minimiser.
- For $f(x) = (x-2)^4$, $x = 2$ is a strict local minimiser.

Isolated Local Minimiser

A point x^* is called an isolated local minimiser if there is a neighbourhood $\mathcal{N}$ of x^* such that x^* is the only local minimiser in $\mathcal{N}$.

Example

$$f(x) = x^4 \cos(1/x) + 2x^4 \quad f(0) = 0$$

- is twice continuously differentiable
- has a strict local minimiser at $x^* = 0$
- however, there are strict local minimisers at many nearby points x_j,
  and $x_j \to 0$ as $j \to \infty$

A Zoom plot of $f(x)$ around $x = 0$

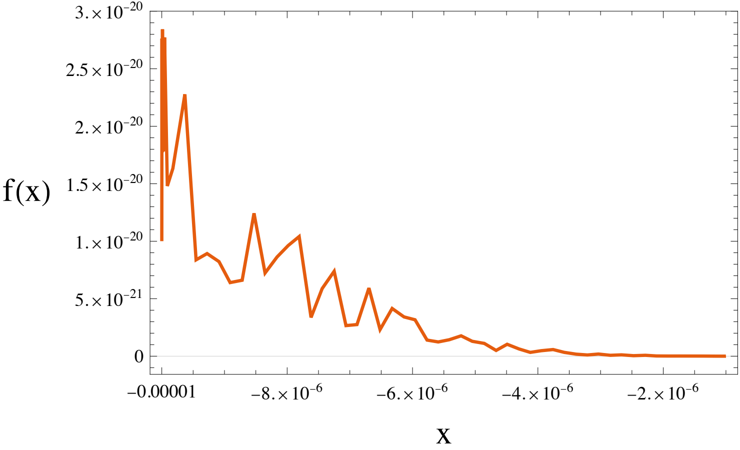

Figure: Showcases many strict local minimisers near $x = 0$.

- Some strict local minimisers are not isolated
- All isolated local minimisers are strict
- It is often difficult to determine a global minimiser for an algorithm, as it often gets trapped in a locality (at a local minimiser)

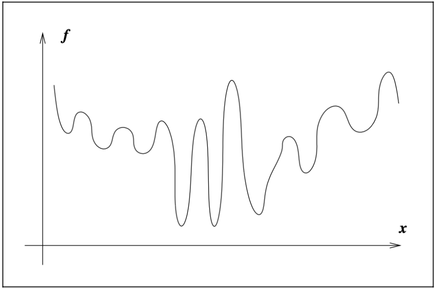

Figure: Showcases a function with many local minimisers.

How to detect minimisers?

- The simplest test being

$$f'(x^*) = 0$$

  is very insufficient to speak anything about the globality of the minimiser.
- These cases (having a lot of local minimisers) is quite standard for optimisation problems.
- Global knowledge about a function f may help identify global minima.
- For convex functions local minimiser is also a global minimiser.

Taylor's Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$. Then we have

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad \text{for some } t \in (0,1)$$

Moreover, if f is twice continuously differentiable, we have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\ p\ dt$$

and

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p, \quad \text{for some } t \in (0,1).$$

Taylor's Theorem Residual Form

Taylor's Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a class of $\mathscr{C}^{k+1}$ on an open convex set $\mathbb{S}$. If $a \in \mathbb{S}$ and $a + h \in \mathbb{S}$, then

$$f(a+h) = \sum_{|\alpha| \leq k} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + R_{a,k}(h)$$

where the remainder is given in Lagrange's form by:

$$R_{a,k}(h) = \sum_{|\alpha|=k+1} \partial^\alpha f(a+ch) \frac{h^\alpha}{\alpha!} \text{ for some } \; c \in (0,1)$$

and in the integral form by

$$R_{a,k}(h) = (k+1) \sum_{|\alpha|=k+1} \frac{h^\alpha}{\alpha!} \int_0^1 (1-t)^k \partial^\alpha f(a+th) dt$$

A bound for the Remainder of Taylor's Theorem

Multi-index Notation

A multi-index is an n-tuple of non-negative integers denoted by (Greek alphabets) $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$

$$
\begin{aligned}
|\alpha| &= \alpha_1 + \alpha_2 + \ldots + \alpha_n \\
\alpha! &= \alpha_1!\alpha! \ldots \alpha_n! \\
x^\alpha &= x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}, \;\; x \in \mathbb{R}^n \\
\partial^\alpha f &= \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_n^{\alpha_n} f = \frac{\partial^{|\alpha|} f}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \ldots \partial_{x_n}^{\alpha_n}}
\end{aligned}
$$

If we know that $|\partial^\alpha f(a + ch)|$ are bounded by some real number M, for $|\alpha| = k + 1$ on the interval $c \in (0, 1)$, then

$$
|R_{a,k}(h)| \leq \frac{M}{(n+1)!} |h|^{k+1}
$$

Recognising A Local Minima

- It seems the only way to conclude a point is a local minimum is by comparing the functional values at every point.
- However, if the function f is smooth, more efficient ways can be thought of to identify local minima.

Theorem (First-Order Necessary Conditions):

If x^* is a local minimizer and f is continuously differentiable in an open neighbourhood of x^*, then

$$\nabla f(x^*) = 0.$$

Remark:

Therefore, for any point to be a minimiser of a function it has to be a critical point.

First-Order Necessary Conditions

Outline of Proof

- By contradiction.
- Let x^* be a minimiser and $\nabla f(x^*) \neq 0$.
- Since $\nabla f(x^*) \neq 0$, let $p = -\nabla f(x^*)$, then
$$p^T \nabla f(x^*) = -||\nabla f(x^*)||^2 < 0$$
- Now, consider
$$g(x) := p^T \nabla f(x) = -(\nabla f(x^*))^T \nabla f(x)$$
$$\implies g(x^*) = -||\nabla f(x^*)||^2$$
- ∇f is continuous near x^*, therefore $g(x)$ is also continuous near x^*.
- $\exists$ a scalar $T > 0$ s.t.
$$g(x^* + tp) < 0 \qquad \text{for all } t \in [0, T]$$
$$\implies p^T \nabla f(x^* + tp) < 0.$$

First-Order Necessary Conditions

- Now for any $\bar{t} \in (0, T]$, we have from the Taylor's theorem

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp), \quad t \in (0, \bar{t})$$

- but,

$$p^T \nabla f(x^* + tp) < 0 \quad \forall\ t \in (0, \bar{t}) \text{ as } \bar{t} \leq T$$
$$\implies f(x^* + \bar{t}p) < f(x^*)\ \forall \bar{t} \in (0, T]$$

- In a neighbourhood of x^* $\exists$ a direction along which a point inside the neighbourhood has a value lesser than at x^* which contradicts the assumption that x^* is a local minimiser.

Stationary Point

Definition

A point x^* is called a stationary point for f if

$$\nabla f(x^*) = 0.$$

- Any local minimiser must be a stationary point for smooth functions.
- B a matrix is positive definite if $p^T Bp > 0$ for all vectors $p \neq 0$.
- positive semi-definite if $p^T Bp \geq 0$ for all p.

Second Order Necessary Conditions

Theorem

If x^* is a local minimiser of f and $\nabla^2 f$ exists and is continuous in an open neighbourhood of x^*, then

$$\nabla f(x^*) = 0 \quad \text{and} \quad \nabla^2 f(x^*) \text{ is positive semi-definite.}$$

Sketch of the Proof

- $\nabla f(x^*) = 0$ from the previous theorem.
- Assume that $\nabla^2 f(x^*)$ is not positive semi-definite.
- Therefore, $\exists$ a vector p s.t.

$$p^T \nabla^2 f(x^*) p < 0$$

Second Order Necessary Conditions

- Now consider the function

$$g(x) = p^T \nabla^2 f(x) p$$

- $g(x^*) < 0$ and since $\nabla^2 f(x)$ is continuous around x^*, $g(x)$ is continuous around x^*
- Therefore $\exists\ T$ s.t. $\forall\ t \in [0, T]$

$$g(x^* + tp) < 0.$$
$$\implies p^T \nabla^2 f(x^* + tp) p < 0.$$

- By doing a Taylor series expansion around x^* we get

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^*) + \frac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp) p$$

$\forall\ \bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$

Second Order Necessary Conditions

1. Therefore,

$$f(x^* + \bar{t}p) = f(x^*) + \frac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp)p$$
$$\implies f(x^* + \bar{p}) < f(x^*)$$

2. which is a contradiction as x^* is a minimiser and in the direction p, the function value is less than that at x^* in any neighbourhood.

Second Order Sufficient Conditions

Theorem

Suppose that $\nabla^2 f$ is continuous in an open neighbourhood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimiser of f.

Sketch of Proof

- Let x^* is not a minimiser.
- For every neighbourhood of x^*, $\exists\ ||\Delta x|| > 0$ s.t.

$$f(x^* + \Delta x) < f(x^*)$$

or $f(x^* + \Delta x) = f(x^*) + \Delta x \nabla f(x^*) + \frac{1}{2}\Delta x^T \nabla^2 f(x^*)\Delta x + R_2(\Delta x)$

- Consider the expression in the R.H.S

$$\frac{1}{2}\Delta x^T \nabla^2 f(x^*)\Delta x + R_2(\Delta x)$$

Second Order Sufficient Conditions

- Define $h(x) = x^T \nabla^2 f(x^*) x$
- Since $\nabla^2 f(x^*)$ is P.D. $h(x) > 0$ for $x \neq 0$.
- Since h is continuous, on the compact set $\{x \mid ||x|| = 1\}$, h should attain its minimum value.
- It has to be > 0. Say it be $\beta > 0$
- Now look at the expression $\Delta x^T \nabla^2 f(x^*) \Delta x$, for $||\Delta x|| > 0$ we can multiply $\frac{1}{||\Delta x||}$ to it and get

$$\frac{\Delta x^T}{||\Delta x||} \nabla^2 f(x^*) \frac{\Delta x}{||\Delta x||} \quad \text{and} \quad \left|\left| \frac{\Delta x}{||\Delta x||} \right|\right| = 1$$

$$\implies \frac{\Delta x^T}{||\Delta x||} \nabla^2 f(x^*) \frac{\Delta x}{||\Delta x||} \geq \beta$$

$$\implies \frac{1}{2} \frac{\Delta x^T}{||\Delta x||} \nabla^2 f(x^*) \frac{\Delta x}{||\Delta x||} \geq \frac{1}{2} \beta$$

- Note that $\lim_{||\Delta x|| \to 0} \frac{R_2(\Delta x)}{||\Delta x||^2} = 0$, one can find a $\delta > 0$ s.t.

$$0 < ||\Delta x|| < \delta \implies \left| \frac{1}{||\Delta x||^2} R_2(\Delta x) \right| < \frac{1}{2}\beta$$

- As a result for all $0 < \Delta x < \delta$ the expression in the R.H.S. ≥ 0.
- Therefore, $f(x^* + \Delta x) > f(x^*)$, which is a contradiction.
- In conclusion x^* is a unique local minimiser.

Remark

The Second order sufficient conditions are not necessary for a point to be a strict local minimiser (without satisfying them as well)
$f(x) = x^4, x^* = 0$ is a local minimiser, but $\nabla^2 f(x^*)$ vanishes, it is not P.D. .

Global Minimiser for Convex Functions

Theorem

When f is convex, any local minimiser x^* is a global minimiser of f. If in addition f is differentiable, then any stationary point x^* is a global minimiser of f.

Sketch of the proof

First Part

- Suppose x^* is a local, but not a global minimiser
- $\exists$ a point $z \in \mathbb{R}^n$ s.t.

$$f(z) < f(x^*)$$

- Consider the line segment that joins x^* to z i.e.

$$x = \lambda z + (1 - \lambda)x^*, \text{ for some } \lambda \in [0, 1]$$

Global Minimiser for Convex Functions

- by convexity of f

  $$f(x) \leq \alpha f(z) + (1-\alpha) f(x^*) < f(x^*) \;\forall\; x \in \mathbb{L}$$

  where $\mathbb{L}$ is the line segment.
- Any neighbourhood of x^* contains a piece of the line segment so there will always be a point $x \in \mathscr{N}$ at which the above inequality is satisfied
- $\implies x^*$ is not a local minimiser.

Global Minimiser for Convex Functions

Second Part

- Suppose x^* is not a global minimiser and choose z as above.

$$\begin{aligned}
\nabla f(x^*)^T(z - x^*) &= \frac{d}{d\lambda} f(x^* + \lambda(z - x^*))|_{\lambda=0} \\
&= \lim_{\lambda \to 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \\
&\le \lim_{\lambda \to 0} \frac{\lambda f(z) + (1 - \lambda) f(x^*) - f(x^*)}{\lambda} \\
&= f(z) - f(x^*) < 0
\end{aligned}$$

$\implies \nabla f(x^*) \neq 0$, or x^* is not a stationary point.