Numerical Optimization

Lecture Notes #2

Unconstrained Optimization; Fundamentals

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Fall 2018

Outline

Last Time

We established that our "favorite problem" for the semester will be of the form

$$\min_{\bar{\mathbf{x}} \in \mathbb{R}^n} f(\bar{\mathbf{x}}),$$

where

$f(\bar{\mathbf{x}})$	*the* **objective function**
$\bar{\mathbf{x}}$	*the vector of variables (a.k.a. unknowns, or parameters.)*

The problem is **unconstrained** since all values of $\bar{\mathbf{x}} \in \mathbb{R}^n$ are allowed.

Further, we established that our initial approach will focus on problems where we do not have any extra factors working against us, *i.e.* we are considering local optimization, continuous variables, and deterministic techniques.

What are we looking for? Global Optimizer

A **solution** to the unconstrained optimization problem is a point $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ such that

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

such a point is called a **global minimizer**.

In order to find a global optimizer we need information about the objective on a global scale.

— Unless we have special information (such as convexity of f), this information is "expensive" since we would have to *evaluate* f in (infinitely?) many points.

What are we looking for? — Local Optimizers, 1 of 3

Most algorithms will take a starting point $\bar{\mathbf{x}}_0$ and use information about f, and possibly its derivative(s) in order to compute a point $\bar{\mathbf{x}}_1$ which is "closer to optimal" than $\bar{\mathbf{x}}_0$, in the sense that

$$f(\bar{\mathbf{x}}_1) \leq f(\bar{\mathbf{x}}_0).$$

Then the algorithm will use information about f + derivative(s) in $\bar{\mathbf{x}}_1$ (and possibly in $\bar{\mathbf{x}}_0$ — this increases the storage requirement) to find $\bar{\mathbf{x}}_2$ such that

$$f(\bar{\mathbf{x}}_2) \leq f(\bar{\mathbf{x}}_1) \leq f(\bar{\mathbf{x}}_0).$$

An algorithm of this type will only be able to find a **local minimizer**.

Fundamentals of Unconstrained Optimization
Optimality
Quick Review...
Characterizing the Solution
Some Fundamental Theorems and Definitions...

What are we looking for? — Local Optimizers, 2 of 3

Definition (Local Minimizer)

A point $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ is a **local minimizer** if there is a neighborhood N of $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ such that $f(\bar{\mathbf{x}}^*) \leq f(\bar{\mathbf{x}})$, $\forall \bar{\mathbf{x}} \in N$.

Note: A neighborhood of $\bar{\mathbf{x}}^*$ is an open set which contains $\bar{\mathbf{x}}^*$.

Note: A local minimizer of this type is sometimes referred to as a **weak local minimizer**. A **strict** or **strong** local minimizer is defined as —

Definition (Strict Local Minimizer)

A point $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ is a **strict local minimizer** if there is a neighborhood N of $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ such that $f(\bar{\mathbf{x}}^*) < f(\bar{\mathbf{x}})$, $\forall \bar{\mathbf{x}} \in N - \{\bar{\mathbf{x}}^*\}$.

Fundamentals of Unconstrained Optimization
Optimality
Quick Review...
Characterizing the Solution
Some Fundamental Theorems and Definitions...

What are we looking for? — Local Optimizers, 3 of 3

Definition (Isolated Local Minimizer)

A point $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ is an **isolated local minimizer** if there is a neighborhood N of $\bar{\mathbf{x}}^* \in \mathbb{R}^n$ such that $\bar{\mathbf{x}}^*$ is the only local minimizer in N.

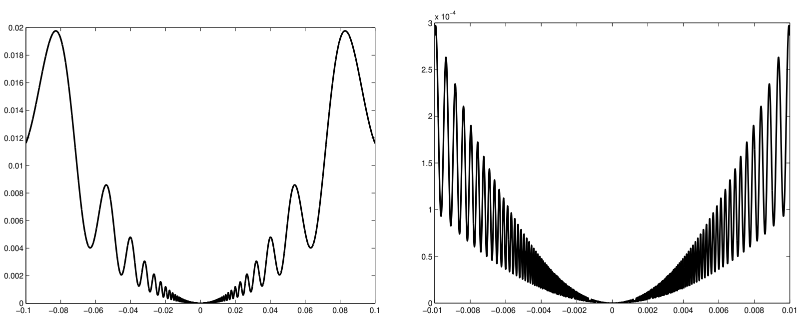

Figure: *The objective $f(x) = x^2(2 + \cos(1/x))$ has a strict local minimizer at $x = 0$, however there are strict local minimizers at infinitely many neighboring points. $x^* = 0$ is not an isolated minimizer.*

Fundamentals of Unconstrained Optimization
Optimality
Quick Review...
Characterizing the Solution
Some Fundamental Theorems and Definitions...

Recognizing A Local Minimum

If we are given a point $\bar{\mathbf{x}} \in \mathbb{R}^n$ how do we know if it is a (local) minimizer??? — Do we have to look at all the points in the neighborhood?

If/when the objective function $f(\bar{\mathbf{x}}) \in \mathbb{R}$ is **differentiable** we can recognize a minimum by looking at the first and second derivatives
— the **gradient** $\nabla f(\bar{\mathbf{x}}) \in \mathbb{R}^n$, and
— the **Hessian*** $\nabla^2 f(\bar{\mathbf{x}}) \in \mathbb{R}^{n \times n}$.

The key tool is the multi-dimensional version of **Taylor's Theorem** (Taylor† expansions/series).

* after Ludwig Otto Hesse (4/22/1811 – 8/4/1874).
† Brook Taylor (8/18/1685 – 12/29/1731).

Illustration: The Gradient (∇f) and the Hessian ($\nabla^2 f$)

Example: Let $\bar{\mathbf{x}} \in \mathbb{R}^3$, *i.e.*

$$\bar{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

then

$$\underbrace{\nabla f(\bar{\mathbf{x}}) = \begin{bmatrix} \frac{\partial f(\bar{\mathbf{x}})}{\partial x_1} \\ \frac{\partial f(\bar{\mathbf{x}})}{\partial x_2} \\ \frac{\partial f(\bar{\mathbf{x}})}{\partial x_3} \end{bmatrix}}_{\text{Gradient}}, \quad \underbrace{\nabla^2 f(\bar{\mathbf{x}}) = \begin{bmatrix} \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_1^2} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_1 \partial x_3} \\ \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_2^2} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_2 \partial x_3} \\ \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_1 \partial x_3} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_2 \partial x_3} & \frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_3^2} \end{bmatrix}}_{\text{Hessian}}.$$

Taylor's Theorem

Theorem (Taylor's Theorem)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable, and that* $\bar{\mathbf{p}} \in \mathbb{R}^n$. *Then,*

$$f(\bar{\mathbf{x}} + \bar{\mathbf{p}}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}} + t\bar{\mathbf{p}})^T \bar{\mathbf{p}},$$

for some $t \in (0,1)$. *Moreover, if* f *is twice continuously differentiable —* $f \in C^2(\mathbb{R}^n)$ *— then*

$$\nabla f(\bar{\mathbf{x}} + \bar{\mathbf{p}}) = \nabla f(\bar{\mathbf{x}}) + \int_0^1 \nabla^2 f(\bar{\mathbf{x}} + t\bar{\mathbf{p}})\bar{\mathbf{p}}\, dt$$

and

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

for some $t \in (0,1)$.

Optimality: First Order Necessary Conditions (Theorem)

Theorem (First-Order Necessary Conditions)

If $\bar{\mathbf{x}}^*$ *is a local minimizer and* f *is continuously differentiable in an open neighborhood of* $\bar{\mathbf{x}}^*$, *then* $\nabla f(\bar{\mathbf{x}}^*) = 0$.

Optimality: First Order Necessary Conditions (Proof)

Proof (By contradiction).

Suppose $\nabla f(\bar{\mathbf{x}}^*) \neq 0$. Let $\bar{\mathbf{p}} = -\nabla f(\bar{\mathbf{x}}^*)$ and realize that $\bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}^*) = -\|\nabla f(\bar{\mathbf{x}}^*)\|^2 < 0$. By continuity of ∇f, there is a scalar $T > 0$ such that

$$\bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}^* + t\bar{\mathbf{p}}) < 0, \quad \forall t \in [0, T]$$

Further, for any $s \in (0, T]$, by Taylor's theorem:

$$f(\bar{\mathbf{x}}^* + s\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}^*) + s\,\underbrace{\bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}^* + t\bar{\mathbf{p}})}_{<0}, \quad \textit{for some } t \in (0, s).$$

Therefore $f(\bar{\mathbf{x}}^* + s\bar{\mathbf{p}}) < f(\bar{\mathbf{x}}^*)$, which contradicts the fact that $\bar{\mathbf{x}}^*$ is a local minimizer. Hence, we must have $\nabla f(\bar{\mathbf{x}}^*) = 0$. □

Optimality: Language and Notation

If $\nabla f(\bar{\mathbf{x}}^*) = 0$, then we call $\bar{\mathbf{x}}^*$ a **stationary point.**

Recall from linear algebra —

Definition (Positive Definite Matrix)

An $n \times n$-matrix A is **Positive Definite** if and only if

$$\forall \bar{\mathbf{x}} \neq 0, \; \bar{\mathbf{x}}^T A \bar{\mathbf{x}} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j > 0.$$

Definition (Positive Semi-Definite Matrix)

An $n \times n$-matrix A is **Positive Semi-Definite** if and only if

$$\forall \bar{\mathbf{x}} \neq 0, \; \bar{\mathbf{x}}^T A \bar{\mathbf{x}} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \geq 0.$$

Optimality: Second-Order Necessary Conditions

Theorem (Second-Order Necessary Conditions)

If $\bar{\mathbf{x}}^*$ is a local minimizer of f and $\nabla^2 f$ is continuous in an open neighborhood of $\bar{\mathbf{x}}^*$, then $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive semi-definite.

Proof.

$\nabla f(\bar{\mathbf{x}}^*) = 0$ follows from the previous proof. We show that $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive semi-definite by contradiction: Assume that $\nabla^2 f(\bar{\mathbf{x}}^*)$ is not positive semi-definite. Then there must exist a vector $\bar{\mathbf{p}}$ such that $\bar{\mathbf{p}}^t \nabla^2 f(\bar{\mathbf{x}}^*)\bar{\mathbf{p}} < 0$. By continuity of $\nabla^2 f$ there is a $T > 0$ such that $\bar{\mathbf{p}}^t \nabla^2 f(\bar{\mathbf{x}}^* + t\bar{\mathbf{p}})\bar{\mathbf{p}} < 0 \; \forall t \in [0, T]$. Now, the Taylor expansion around $\bar{\mathbf{x}}^*$, shows that $\forall s \in (0, T]$ there exists $t \in (0, T)$ such that

$$f(\bar{\mathbf{x}}^* + s\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}^*) + s\bar{\mathbf{p}}^T \underbrace{\nabla f(\bar{\mathbf{x}}^*)}_{=0} + \frac{1}{2}s^2 \underbrace{\bar{\mathbf{p}}^T \nabla^2 f(\bar{\mathbf{x}}^* + t\bar{\mathbf{p}})\bar{\mathbf{p}}}_{<0}.$$

Hence $f(\bar{\mathbf{x}}^* + s\bar{\mathbf{p}}) < f(\bar{\mathbf{x}}^*)$, which is a contradiction. □

Optimality: Necessary vs. Sufficient Conditions

The conditions we have outlined so far are **necessary**; hence **if** $\bar{\mathbf{x}}^*$ is a minimum, **then** the conditions must hold.

It is more useful to have a set of **sufficient conditions**, so that **if** the conditions are satisfied (at $\bar{\mathbf{x}}^*$), **then** $\bar{\mathbf{x}}^*$ is a minimum.

The **second order sufficient conditions** guarantee that $\bar{\mathbf{x}}^*$ is a strict local minimizer of f, and the **convexity** of f guarantees that any local minimizer is a global minimizer...

Optimality: Second-order Sufficient Conditions (Theorem)

Theorem (Second-Order Sufficient Conditions)

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of $\bar{\mathbf{x}}^*$ and that $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive definite. Then $\bar{\mathbf{x}}^*$ is a strict local minimizer of f.

Optimality: Second-order Sufficient Conditions (Proof)

Proof.

Since the Hessian $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive definite, we can find a open ball of positive radius r, $D(r;\bar{\mathbf{x}}^*) = \{\bar{\mathbf{y}} \in \mathbb{R}^n : \|\bar{\mathbf{x}}^* - \bar{\mathbf{y}}\| < r\}$, so that $\nabla^2 f(\bar{\mathbf{y}})$ is positive definite $\forall \bar{\mathbf{y}} \in D$. Now, for any vector $\bar{\mathbf{p}}$ such that $\|\bar{\mathbf{p}}\| < r$, we have $\bar{\mathbf{x}}^* + \bar{\mathbf{p}} \in D$ and therefore (by Taylor)

$$f(\bar{\mathbf{x}}^* + \bar{\mathbf{p}}) = f(\bar{\mathbf{x}}^*) + \bar{\mathbf{p}}^T \underbrace{\nabla f(\bar{\mathbf{x}}^*)}_{=0} + \frac{1}{2} \underbrace{\bar{\mathbf{p}}^T \nabla^2 f(\bar{\mathbf{x}}^* + t\bar{\mathbf{p}})\bar{\mathbf{p}}}_{>0}$$

for some $t \in (0,1)$. Hence it follows that $f(\bar{\mathbf{x}}^*) < f(\bar{\mathbf{x}}^* + \bar{\mathbf{p}})$, and so $\bar{\mathbf{x}}^*$ must be a strict local minimizer. □

Optimality: Convexity

Theorem

When the objective function f is **convex**, *any local minimizer $\bar{\mathbf{x}}^*$ is also a global minimizer of f. If in addition f is differentiable, then any stationary point $\bar{\mathbf{x}}^*$ is a global minimizer of f.*

Optimality: Convexity

Proof (part-1).

Suppose that $\bar{\mathbf{x}}^*$ is a local, but not a global minimizer. Then there must exist a point $\bar{\mathbf{z}} \in \mathbb{R}^n$ such that $f(\bar{\mathbf{z}}) < f(\bar{\mathbf{x}}^*)$. Consider the line-segment that joins $\bar{\mathbf{x}}^*$ and $\bar{\mathbf{z}}$:

$$\bar{\mathbf{y}}(\lambda) = \lambda\bar{\mathbf{z}} + (1-\lambda)\bar{\mathbf{x}}^*, \quad \lambda \in [0,1]$$

Since f is convex we must have **[by definition]**

$$f(\bar{\mathbf{y}}(\lambda)) \le \lambda f(\bar{\mathbf{z}}) + (1-\lambda) f(\bar{\mathbf{x}}^*) < f(\bar{\mathbf{x}}^*), \quad \lambda \in (0,1]$$

Every neighborhood of $\bar{\mathbf{x}}^*$ will contain a piece of the line-segment, hence $\bar{\mathbf{x}}^*$ cannot be a local minimizer. □

Optimality: Convexity

Proof (part-2).

Suppose that $\bar{\mathbf{x}}^*$ is a local but not a global minimizer, and let $\bar{\mathbf{z}}$ be such that $f(\bar{\mathbf{z}}) < f(\bar{\mathbf{x}}^*)$. Using convexity, and the definition of a directional derivative (NW[2nd] p-628), we have

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

Therefore, $\nabla f(\bar{\mathbf{x}}^*) \neq 0$, so $\bar{\mathbf{x}}^*$ cannot be a stationary point. This contradicts the supposition that f is a local minimum. □

Optimality: Theorems and Algorithms

The theorems we have shown — all of which are based on elementary (vector) calculus — are the backbone of unconstrained optimization algorithms.

Since we usually do not have a global understanding of f, the algorithms will seek stationary points, *i.e.* solve the problem

$$\nabla \mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}.$$

When $\bar{\mathbf{x}} \in \mathbb{R}^n$, this is a system of n (generally) non-linear equations.

Hence, there is a strong connection between the solution of non-linear equations and unconstrained optimization.

— We will focus on developing an optimization framework, and in the last few weeks of the semester we will use it to solve non-linear equations.

Algorithms — An Overview

The algorithms we study start with an initial (sub-optimal) guess $\bar{\mathbf{x}}_0$, and generate a sequence of iterates $\{\bar{\mathbf{x}}_k\}_{k=1,\dots,N}$.

The sequence is terminated when either

[success] We have approximated a solution up to desired accuracy.

[failure] No more progress can be made.

Different algorithms make different decisions in how to move from $\bar{\mathbf{x}}_k$ to the next iterate $\bar{\mathbf{x}}_{k+1}$.

Many algorithms are **monotone**, *i.e.* $f(\bar{\mathbf{x}}_{k+1}) < f(\bar{\mathbf{x}}_k)$, $\forall k \geq 0$, but there exist **non-monotone** algorithms. Even a non-monotone algorithm is required to *eventually* decrease — how else can we reach a minimum? Typically $f(\bar{\mathbf{x}}_{k+m}) < f(\bar{\mathbf{x}}_k)$ is required for some fixed value $m > 0$ and $\forall k \geq 0$.

Moving from $\bar{\mathbf{x}}_k$ to $\bar{\mathbf{x}}_{k+1}$ — Line Search

Most optimization algorithms use one of two fundamental strategies for finding the next iterate: —

1. Line search based algorithms reduce the n-dimensional optimization problem

$$\min_{\bar{\mathbf{x}} \in \mathbb{R}^n} f(\bar{\mathbf{x}}),$$

with a one-dimensional problem:

$$\min_{\alpha > 0} f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k),$$

where $\bar{\mathbf{p}}_k$ is a chosen **search direction**. Clearly, how cleverly we select $\bar{\mathbf{p}}_k$ will affect how much progress we can make in each iteration.

— The intuitive choice gives a slow scheme!

Moving from $\bar{\mathbf{x}}_k$ to $\bar{\mathbf{x}}_{k+1}$ — Trust Region, 1 of 2

2. Trust region based methods take a completely different approach. — Using information gathered about the objective f, *i.e.* function values, gradients, Hessians, etc. during the iteration, a simpler **model function** is generated.

A good model function $m_k(\bar{\mathbf{x}})$ approximates the behavior of $f(\bar{\mathbf{x}})$ in a neighborhood of $\bar{\mathbf{x}}_k$, *e.g.* Taylor expansion

$$m_k(\bar{\mathbf{x}}_k + \bar{\mathbf{p}}) = f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2} \bar{\mathbf{p}}^T H_k \bar{\mathbf{p}},$$

where H_k is the full Hessian $\nabla^2 f(\bar{\mathbf{x}}_k)$ (expensive) or a clever approximation thereof.

Moving from $\bar{\mathbf{x}}_k$ to $\bar{\mathbf{x}}_{k+1}$ — Trust Region, 2 of 2

The model is chosen simple enough that the optimization problem

$$\min_{p \in N(\bar{\mathbf{x}}_k)} m_k(\bar{\mathbf{x}}_k + \bar{\mathbf{p}}),$$

can be solved quickly. The neighborhood $N(\bar{\mathbf{x}}_k)$ of $\bar{\mathbf{x}}_k$ specifies the region in which we trust the model.

A simple model can only capture the local behavior of f — think about how the Taylor expansion approximates a function well close to the expansion point, but not very well further away.

Usually the trust region is a ball in $\mathbb{R}^n$, *i.e.*

$$N(\bar{\mathbf{x}}_k) = \{\bar{\mathbf{p}} : \|\bar{\mathbf{p}} - \bar{\mathbf{x}}_k\| \leq r\},$$

but elliptical or box-shaped trust regions are sometimes used.

Fundamentals of Unconstrained Optimization
Optimality
Necessary vs. Sufficient Conditions; Convexity
From Theorems to Algorithms...

Line Search vs. Trust Region

Step	Line Search	Trust Region
1	*Choose a search direction $\bar{\mathbf{p}}_k$.*	*Establish the maximum distance — the size of the trust region.*
2	*Identify the distance, e.g. the step length in the search direction.*	*Find the direction in the trust region.*

Table: *Line search and trust region methods handle the selection of direction and distance in opposite order.*

Next time:

— **Rate of Convergence**.

— Line search methods, detailed discussion.

Fundamentals of Unconstrained Optimization
Optimality
Necessary vs. Sufficient Conditions; Convexity
From Theorems to Algorithms...

Index