10.007: Modeling the Systems World

Lecture 2

Unconstrained Optimization

Term 3, 2017

Singapore University of
Technology & Design

Overview

Unconstrained optimization:

- Definition of global and local optima.
- Necessary conditions that local optima must satisfy.
- Sufficient conditions that guarantee that a given point is a local minimum.
- Unboundedness

Global vs. local

Suppose we have the unconstrained optimization problem:

$$\min\{f(x) : x \in \mathbb{R}^n\} \qquad \text{(P)}.$$

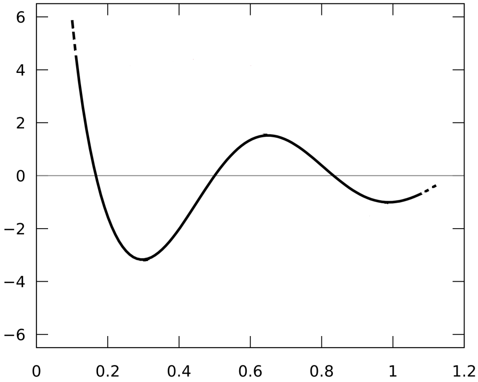
Definition

The point x^* is a global minimum of (P) if $f(x^*) \leqslant f(x)\ \forall x \in \mathbb{R}^n$.

The point x^* is a local minimum of (P) if there exists $\varepsilon > 0$ (possibly very small) such that $f(x^*) \leqslant f(x)$ for all points $x \in \mathbb{R}^n$ such that $\|x^* - x\| \leqslant \varepsilon$.

Global vs. local

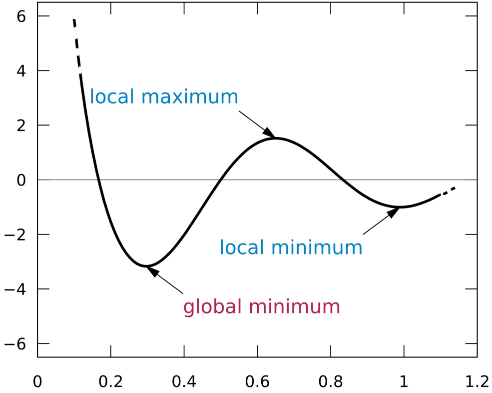
Which of these points are global/local minima? Which are global/local maxima?



When is a local minimum automatically a global one?

Global vs. local

Which of these points are global/local minima? Which are global/local maxima?



When is a local minimum automatically a global one?

A blast from the past

We first focus on unconstrained problems and discuss how to find the local minima of a problem given as:

$$\min\{f(x) : x \in \mathbb{R}^n\} \qquad \text{(P)}.$$

Single-variable function $f(x)$:

- Find points with $f'(x) = 0$.

  For all such points:
- If $f''(x) > 0$, x is local min.
- If $f''(x) = 0$, check further.
- If $f''(x) < 0$, x is local max.

Two-variable function $f(x, y)$:

- Find points with $\nabla f(x, y) = (0, 0)$.

  For all such points:
- If $\det H_f(x, y) > 0$ and $f_{xx}(x, y) > 0$, (x, y) is local min. ($f_{xx}(x, y) < 0$, (x, y) is local max.)
- If $\det H_f(x, y) = 0$, check further.
- If $\det H_f(x, y) < 0$, saddle point.

A blast from the past

You have seen why these methods work for functions with one or two variables in the Math 2 class. (Please go back and check your notes: especially, Cohort 18 of Math 2!)

Today (and next week) we will generalize them to functions with n variables and understand them better.

To understand optimization in higher dimensions, we need to extend our toolbox. So, we will next learn how to approximate multivariable functions using partial derivatives.

Taylor series

Taylor expansion at x^* up to first-order:

- One-variable, $x^* \in \mathbb{R}$, $\alpha \in \mathbb{R}$:

$$f(x^* + \alpha) = f(x^*) + \alpha f'(x^*) + r(\alpha)$$

- Multi-variable, $x^*, u \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha)$$

Here, we assume that $\|u\| = 1$: u is a unit vector that represents the direction of displacement from x^*.

Also, $r(\alpha)$ is a constant that is much smaller than α when α is small enough.

Taylor series

Similarly, we can find the Taylor expansion at x^* up to second-order:

- One-variable, $x^* \in \mathbb{R}$, $\alpha \in \mathbb{R}$:

$$f(x^* + \alpha) = f(x^*) + \alpha f'(x^*) + \frac{\alpha^2}{2} f''(x^*) + r(\alpha^2)$$

- Multi-variable, $x^*, u \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$:

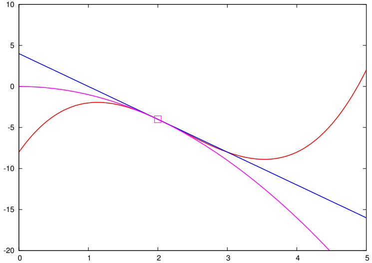
$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + \frac{\alpha^2}{2} u^T H_f(x^*) u + r(\alpha^2)$$

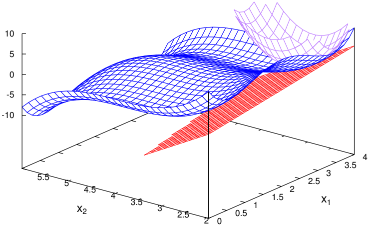
Here, we assume that $\|u\| = 1$: u is a unit vector that represents the direction of displacement from x^*.

In general, the Taylor series up to k-th order has an error term $r(\alpha^k)$ with the property that $\lim_{\alpha \to 0} r(\alpha^k)/\alpha^k = 0$.

Taylor series

The higher the better “locally"...





A note on necessary versus sufficient!

It is necessary to have clouds, when it is raining.

But having clouds is NOT sufficient to have rain.

A note on necessary versus sufficient!

What is a necessary condition to win a football game?

What is a sufficient condition to win a football game?

First-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

First-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

Why are the first-order necessary conditions useful ?

First-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

Why are the first-order necessary conditions useful ?

Single-variable function $f(x)$:

- Find points with $f'(x) = 0$.

Two-variable function $f(x, y)$:

- Find points with $\nabla f(x, y) = (0, 0)$.

Reason: if the minimum exists, it must be one of the points such that $\nabla f(x^*) = \vec{0}$. These are called the stationary points.

Second-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. If x^* is an unconstrained local minimum of f, then $H_f(x^*)$ is positive semidefinite.

Second-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. If x^* is an unconstrained local minimum of f, then $H_f(x^*)$ is positive semidefinite.

Why are the second-order necessary conditions useful ?

Second-order necessary conditions

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. If x^* is an unconstrained local minimum of f, then $H_f(x^*)$ is positive semidefinite.

Why are the second-order necessary conditions useful ?

Single-variable function $f(x)$:

- Find points with $f'(x) = 0$.
- For all such points: check $f''(x) \geqslant 0$. If not, it is not a local minimum!

Two-variable function $f(x, y)$:

- Find points with $\nabla f(x, y) = (0, 0)$.
- For all such points: check $H_f(x, y)$ is psd. If not, it is not a local minimum!.

Reason: We can eliminate the stationary points that fail to satisfy this condition. These are saddle points.

Positive (semi)definite matrices

Definition

A symmetric matrix $H \in \mathbb{R}^{n\times n}$ is positive semidefinite (psd) if $x^T Hx \geqslant 0$ for all vectors $x \in \mathbb{R}^n$. It is positive definite (pd) if $x^T Hx > 0$ for all nonzero vectors $x \in \mathbb{R}^n$. It is indefinite if $x^T Hx$ could be positive or negative depending on x.

Equivalent definition

A symmetric matrix $H \in \mathbb{R}^{n\times n}$ is positive semidefinite if all its eigenvalues are $\geqslant 0$. It is positive definite if all its eigenvalues are > 0. It is indefinite if it has both positive and negative eigenvalues.

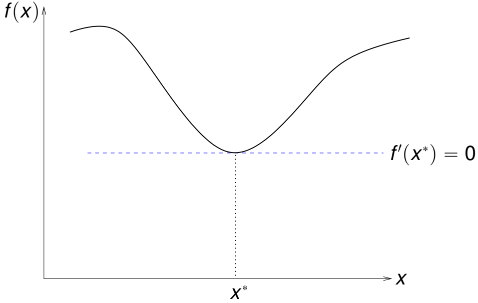
Can you show that these definitions are equivalent using the "definition of the eigenvalues" (one direction) and "diagonalization of symmetric matrices" (reverse direction) that you have learned in Math 2? Try it out!

Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

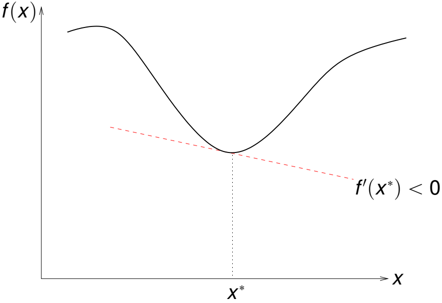


Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

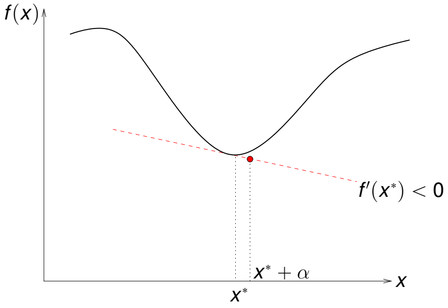


Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

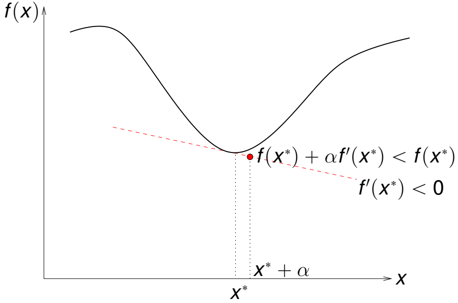


Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

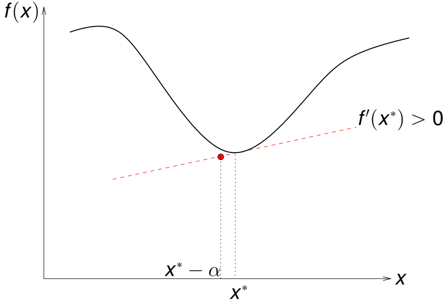


Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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

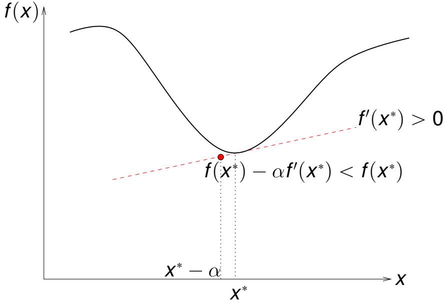


Proof of the first proposition (1D)

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If x^* is an unconstrained local minimum of f, then

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



A simple optimization procedure?

A simple optimization procedure? NOT YET!!!

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.

Based on the two propositions above, can we use this method to find the minimum of f?

1. Find all points $z \in \mathbb{R}^n$ such that $\nabla f(z) = \vec{0}$.
2. Among those, find all points where $H_f(z)$ is psd.
3. Take the point with the smallest function value.

A simple optimization procedure? NOT YET!!!

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.

Based on the two propositions above, can we use this method to find the minimum of f?

1. Find all points $z \in \mathbb{R}^n$ such that $\nabla f(z) = \vec{0}$.
2. Among those, find all points where $H_f(z)$ is psd.
3. Take the point with the smallest function value.

If there are minima, they will all be contained in the set of points remaining after step 2. This does not mean that there is always a minimum: the method above may fail.

Why this method fails

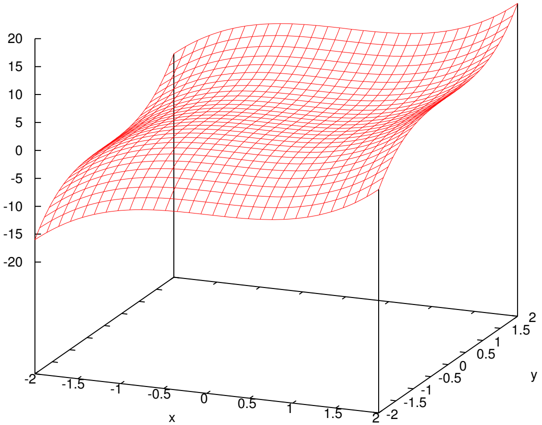
Consider the function $f(x, y) = x^3 + y^3$.

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 \\ 3y^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

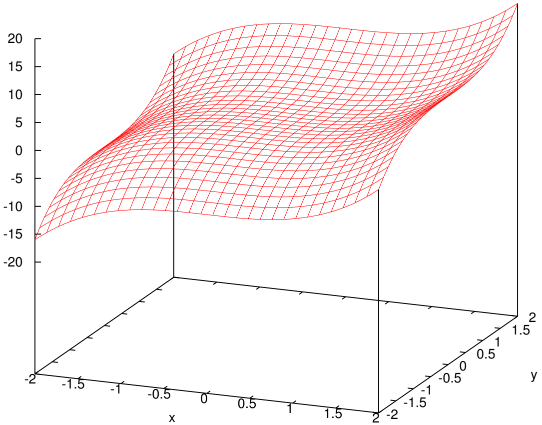
$$H_f(x, y) = \begin{pmatrix} 6x & 0 \\ 0 & 6y \end{pmatrix}, \qquad \text{and } H_f(0, 0) \text{ is psd.}$$

The point $(0, 0)$ satisfies both necessary conditions. But it is not a minimum!

Why this method fails



Why this method fails



The trouble here is that the function has no minimum over $\mathbb{R}^2$.

Question: do we have some way of verifying that we found a local minimum?

A sufficient condition

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Suppose $x^* \in S$ satisfies:

$$\nabla f(x^*) = \vec{0}, \qquad H_f(x^*) \text{ is pd}.$$

Then x^* is a (strict) unconstrained local minimum of f.

A sufficient condition

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Suppose $x^* \in S$ satisfies:

$$\nabla f(x^*) = \vec{0}, \qquad H_f(x^*) \text{ is pd.}$$

Then x^* is a (strict) unconstrained local minimum of f.

Single-variable function $f(x)$:

- Find points with $f'(x) = 0$.
- For all such points: check $f''(x) > 0$.

Two-variable function $f(x, y)$:

- Find points with $\nabla f(x, y) = (0, 0)$.
- For all such points: check $H_f(x, y)$ is pd.

A simple optimization procedure?

almost there, hang on little smurfs...

A simple optimization procedure

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.

Based on the three propositions above, we have a method to find the minimum of f .

1. Find all points $z \in \mathbb{R}^n$ such that $\nabla f(z) = \vec{0}$.
2. Among those, find all points where $H_f(z)$ is psd.
    2.1 Among those, all points where $H_f(z)$ is pd are local minima.
    2.2 Further investigate points where $H_f(z)$ is psd and $\det H_f(z) = 0$.

A simple optimization procedure

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.

Based on the three propositions above, we have a method to find the minimum of f .

1. Find all points $z \in \mathbb{R}^n$ such that $\nabla f(z) = \vec{0}$.
2. Among those, find all points where $H_f(z)$ is psd.
    2.1 Among those, all points where $H_f(z)$ is pd are local minima.
    2.2 Further investigate points where $H_f(z)$ is psd and $\det H_f(z) = 0$.

If there are local minima, they will be identified in Step 2. If there are no points left after Step 2, there is neither local nor global minimum.

The problem is either unbounded or the minimum is not attained!

Unboundedness

Let $f(x, y) = x + y + 2$. Note that,

$\lim_{N\to\infty} f(-N, -N) = \lim_{N\to\infty} -2N + 2 = -\infty$.

Therefore, this function is unbounded (below and above) and has neither minimum nor maximum.

Unboundedness

Let $f(x, y) = x + y + 2$. Note that,

$\lim_{N\to\infty} f(-N, -N) = \lim_{N\to\infty} -2N + 2 = -\infty$.

Therefore, this function is unbounded (below and above) and has neither minimum nor maximum.

Self study: Consider $-x^2 + xy - y^2$, $-e^{(x+y)}$.

Are they unbounded below/above?

Do they attain a minimum/maximum?

An example

Find the global minimum of the following problem:

$$\min x^2 + 3x + y^2 - 2y \quad \text{s.t. } x, y \in \mathbb{R}$$

Let $f(x, y)$ be the objective function, and set $\nabla f = 0$:

$$\frac{\partial f}{\partial x} = 2x + 3 = 0, \qquad \frac{\partial f}{\partial y} = 2y - 2 = 0.$$

Solution: $x = -3/2, y = 1$. Compute the Hessian at $(-3/2, 1)$:

$$H_f(-3/2, 1) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \det(H_f) = 4 > 0.$$

Local minimum at $(-3/2, 1)$. Is this a global minimum?

Wrap-up

We now have both necessary and sufficient conditions for local optima of an unconstrained problem. The sufficient conditions are more restrictive.

We can use the necessary conditions to find candidate local optima. We can use the sufficient conditions to find guaranteed local optima.

The conditions studied today can be used to develop general solution methods for these problems. They can also be used to solve small problems by hand.

When the problem size is large, we can find points that satisfy these conditions by specialized algorithms. We will learn a popular method, the steepest descent algorithm, next week.

Summary

- Presented necessary conditions for optimality:
  - First-order,
  - Second-order.
- Presented a sufficient condition for optimality.
- Understood where the difference lies.
- Revisited our optimization toolbox from Math I & II.

Proof of the first-order necessary conditions

Recall the first-order Taylor expansion at x^*:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha).$$

Proof of the first-order necessary conditions

Recall the first-order Taylor expansion at x^*:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha).$$

If $\nabla f(x^*) \neq 0$, then there must be a vector u such that $\|u\| = 1$ and $\nabla f(x^*)^T u < 0$.

Proof of the first-order necessary conditions

Recall the first-order Taylor expansion at x^*:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha).$$

If $\nabla f(x^*) \neq 0$, then there must be a vector u such that $\|u\| = 1$ and $\nabla f(x^*)^T u < 0$. By definition of local optimum, there exists $\varepsilon > 0$ such that for all $\alpha \leqslant \varepsilon$ we have $f(x^* + \alpha u) \geqslant f(x^*)$, so for small α:

$$0 \leqslant \frac{f(x^* + \alpha u) - f(x^*)}{\alpha} = \nabla f(x^*)^T u + \frac{r(\alpha)}{\alpha}.$$

Proof of the first-order necessary conditions

Recall the first-order Taylor expansion at x^*:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha).$$

If $\nabla f(x^*) \neq 0$, then there must be a vector u such that $\|u\| = 1$ and $\nabla f(x^*)^T u < 0$. By definition of local optimum, there exists $\varepsilon > 0$ such that for all $\alpha \leqslant \varepsilon$ we have $f(x^* + \alpha u) \geqslant f(x^*)$, so for small α:

$$0 \leqslant \frac{f(x^* + \alpha u) - f(x^*)}{\alpha} = \nabla f(x^*)^T u + \frac{r(\alpha)}{\alpha}.$$

Take the limit for $\alpha \to 0$:

$$\begin{aligned} 0 &\leqslant \lim_{\alpha \to 0} \frac{f(x^* + \alpha u) - f(x^*)}{\alpha} = \nabla f(x^*)^T u + \lim_{\alpha \to 0} \frac{r(\alpha)}{\alpha} \\ &= \nabla f(x^*)^T u < 0. \end{aligned}$$

This is impossible, therefore we must have $\nabla f(x^*) = \vec{0}$.

Proof of the first-order necessary conditions

Recall the first-order Taylor expansion at x^*:

$$f(x^* + \alpha u) = f(x^*) + \alpha \nabla f(x^*)^T u + r(\alpha).$$

If $\nabla f(x^*) \neq 0$, then there must be a vector u such that $\|u\| = 1$ and $\nabla f(x^*)^T u < 0$. By definition of local optimum, there exists $\varepsilon > 0$ such that for all $\alpha \leqslant \varepsilon$ we have $f(x^* + \alpha u) \geqslant f(x^*)$, so for small α:

$$0 \leqslant \frac{f(x^* + \alpha u) - f(x^*)}{\alpha} = \nabla f(x^*)^T u + \frac{r(\alpha)}{\alpha}.$$

Take the limit for $\alpha \to 0$:

$$\begin{aligned} 0 &\leqslant \lim_{\alpha \to 0} \frac{f(x^* + \alpha u) - f(x^*)}{\alpha} = \nabla f(x^*)^T u + \lim_{\alpha \to 0} \frac{r(\alpha)}{\alpha} \\ &= \nabla f(x^*)^T u < 0. \end{aligned}$$

This is impossible, therefore we must have $\nabla f(x^*) = \vec{0}$.
This is a proof by contradiction!

Proof of the second proposition

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. If x^* is an unconstrained local minimum of f, then $H_f(x^*)$ is positive semidefinite.

Proof of the second proposition

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. If x^* is an unconstrained local minimum of f, then $H_f(x^*)$ is positive semidefinite.

The proof is similar to the first condition and is not too hard. It also gives insight on why the Hessian is so important.

You will learn more about Taylor series and the details of this proof in the cohorts!

A note on sufficiency conditions

Note that when the second order sufficiency conditions are satisfied, we have a strict minimum. The definition is as follows:

$$\min\{f(x) : x \in \mathbb{R}^n\} \qquad \text{(P)}.$$

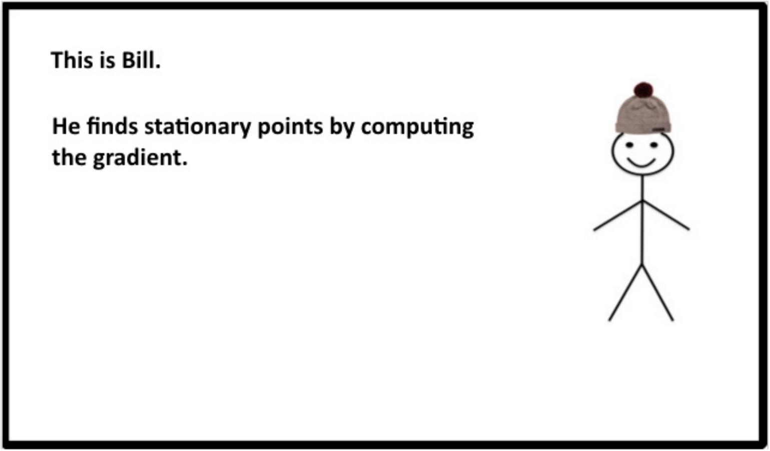
Definition

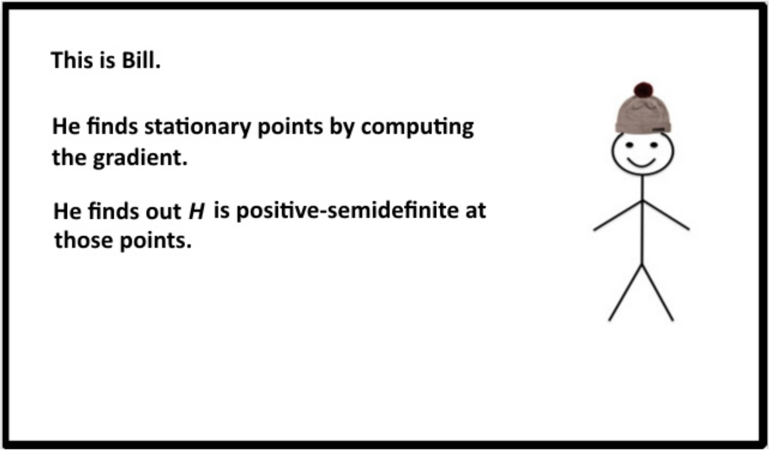
The point x^* is a strict global minimum of (P) if

$$f(x^*) < f(x) \; \forall x \in \mathbb{R}^n.$$

The point x^* is a strict local minimum of (P) if there exists $\varepsilon > 0$ (possibly very small) such that $f(x^*) < f(x)$ for all points $x \in \mathbb{R}^n$ such that $\|x^* - x\| \leqslant \varepsilon$.


This is Bill.


This is Bill.
He finds stationary points by computing the gradient.


This is Bill.
He finds stationary points by computing the gradient.
He finds out H is positive-semidefinite at those points.

This is Bill.

He finds stationary points by computing the gradient.

He finds out H is positive-semidefinite at those points.

He knows it's a necessary condition.

This is Bill.

He finds stationary points by computing the gradient.

He finds out H is positive-semidefinite at those points.

He knows it's a necessary condition.

Bill is smart.

This is Bill.

He finds stationary points by computing the gradient.

He finds out H is positive-semidefinite at those points.

He knows it's a necessary condition.

Bill is smart.

Be like Bill.