AMS 761 Nonlinear Optimization I	**FALL 2023**

Lecture 2: August 31

Lecturer: Mateo Díaz — *Scribe: Emmett Springer*

2.1 Calculus review

2.1.1 Gradient

Consider a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ where there exists a gradient vector field $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$. This gradient satisfies:

$$\lim_{\bar{u} \to \bar{0}} \frac{f(\bar{x} + \bar{u}) - (f(\bar{x}) + \langle \nabla f(\bar{x}), \bar{u} \rangle)}{\| \bar{u} \|} = 0.$$

We can equivalently define the gradient as:

$$\nabla f(\bar{x}) = \begin{bmatrix} \frac{\delta f(\bar{x})}{\delta x_1} \\ \vdots \\ \frac{\delta f(\bar{x})}{\delta x_n} \end{bmatrix} \tag{2.1}$$

The intuition for this is illustrated with the graph below. When $\bar{u}$ is very small, it will look like the slope of the first order approximation almost matches that of the function at x. Note that the gradient will always give you a local approximation, not a global approximation.

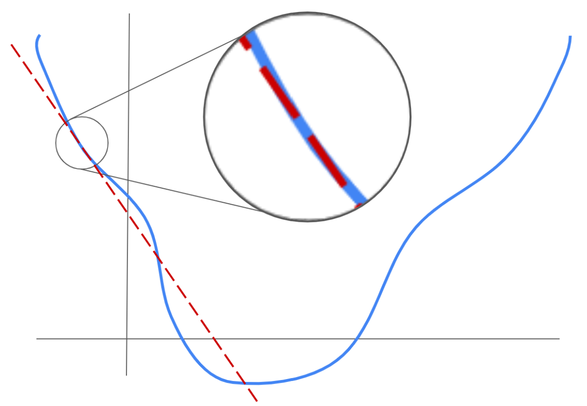

Figure 2.1: Intuition for the gradient. As $\bar{u} \to \bar{0}$, it aligns with $f(\bar{x})$.

We say that a function f is twice differential (C^2) if an operator $\nabla^2 f(x) : \mathbb{R}^d \to \mathbb{R}^d$ exists $\forall x$ s.t.

$$\lim_{\bar{u} \to \bar{0}} \frac{f(\bar{x} + \bar{u}) - (f(\bar{x}) + \nabla f(\bar{x})^T \bar{u} + (\frac{\bar{u}^T \nabla^2 f(\bar{x}) \bar{u}}{2})}{\| \bar{u} \|^2} = 0.$$

This is a more complicated expansion beyond the gradient, but it has a similar intuition. If we want to see how this function looks as we get close to $\bar{u}$, we will be a quadratic function rather than a linear function. Here we get rid of all things size $\| \bar{u} \|^2$ as opposed to getting rid of things size $\| \bar{u} \|$ as with the previous definition. This makes it a good approximation up to second order terms as opposed to first order terms.

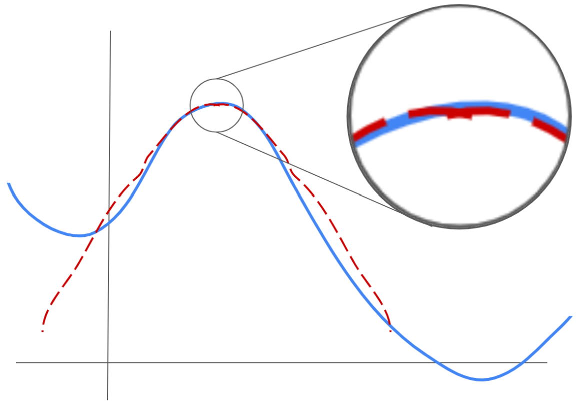

Figure 2.2: Intuition for $\nabla^2 f(x) : \mathbb{R}^d \to \mathbb{R}^d$.

2.1.2 Chain rule

Theorem 2.1 *Let f be a twice-differential function. Pick $\bar{x}, \bar{u} \in \mathbb{R}^d$, and define $\varphi : \mathbb{R} \to \mathbb{R}$, given by $\rho(t) = f(\bar{x} + \bar{u})$. Then:*

- $\varphi'(t) = \nabla f(\bar{x} + t\bar{u})^T \bar{u}$.
- $\varphi''(t) = \bar{u}^T \nabla^2 f(\bar{x} + t\bar{u})\bar{u}$.

The application of chain rule is important to nonlinear optimization, because in many scenarios we only need to analyze one direction of the function at a time.

2.1.3 First order Taylor approximation

Theorem 2.2 *Let f be a function with an L-Lipschitz gradient.*

(Note that f has an L-Lipscitz gradient if for some $L > 0$, $\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|, \forall x, y$)

Then $\forall \bar{x}, \bar{u} \in \mathbb{R}^d$:

$$| f(\bar{x} + \bar{u}) - (f(\bar{x}) + \langle \nabla f(\bar{x}), u \rangle) | \leq \frac{L}{2} \| u \|^2 .$$

This tells us (1) that the limit is going to 0 and (2) the rate at which the limit is going to zero.

2.1.4 Second order Taylor approximation

Theorem 2.3 *Let f be a twice differentiable set with $\nabla^2 f$ is Q-Lipschitz. Then:*

$$| f(\bar{x}+\bar{u}) - (f(\bar{x}) + \langle \nabla f(\bar{x}), \bar{u}\rangle + \frac{1}{2}\bar{u}^T \nabla^2 f(\bar{x})\bar{u}) | \leq \frac{Q}{6} \parallel \bar{u} \parallel^3 .$$

The second-order Taylor approximation is a quadratic approximation, as opposed to a linear approximation with the first-order.

2.2 Optimality conditions

2.2.1 Global vs local minimizers

When solving for $\min_{\bar{x}\in\mathbb{R}} f(\bar{x})$, we may encounter different types of global and local minimizers:

- A point $\bar{x}^*$ is a global minimizer if $\forall \bar{u} \in \mathbb{R}^d, f(\bar{x}^*) \leq f(\bar{u})$.
- A point $\bar{x}^*$ is a local minimizer if there exists a small ball around it, B, s.t. $\bar{u} \in B_\epsilon(\bar{x}^*)$ $f(\bar{x}^*) \leq f(\bar{u})$.
- A point $\bar{x}^*$ is a strict local minimizer if there exists ϵ s.t. $\forall \bar{u} \in B_\epsilon(\bar{x}^*)$ $f(\bar{x}^*) < f(\bar{u})$.

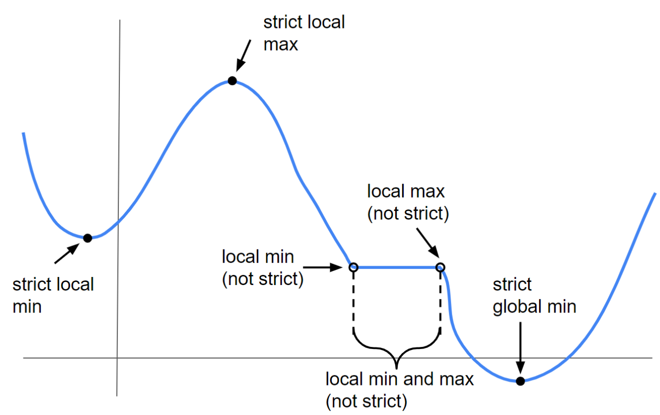

Figure 2.3: Example of three types of minimizers (and maximizers).

When optimizing, we will not always reach an exactly optimal solution. Therefore, it is useful to utilize **optimality conditions**, which give us a way to check whether a point is a minimizer, or at least satisfies some conditions that minimizers satisfy: There are four types of minimizers we'll discuss:

- First order necessary condition.
- First order sufficient condition for convex functions.
- Second order necessary condition.
- Second order sufficient condition.

2.2.2 First order necessary optimality condition

Any global or local minimizer must satisfy the 1st order necessary optimality condition, but its fulfillment does not guarantee a global or local minimizer. Thus making the condition necessary, but not sufficient. Assume a function f is differentiable. Then if $\bar{x}^*$ is a minimizer, it satisfies that $\nabla f(\bar{x}^*) = 0$.

The intuition for this definition is illustrated below. If the tangent to the point $\bar{x}^*$ is non-zero, then there must be a direction in which the function goes "downhill", meaning that there exists a better point.

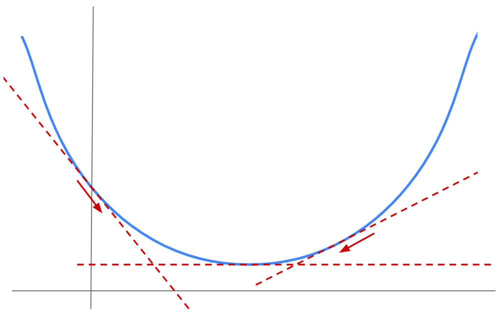

Figure 2.4: Intuition for the first-order necessary optimality condition.

Theorem 2.4 *If $\bar{x}^*$ is a minimizer, then $\nabla f(\bar{x}^*) = 0$.*

Proof: Seeking contradition, assume that $\nabla f(\bar{x}^*) \neq 0$.

Take $u = \frac{-\nabla f(\bar{x}^*)}{\|\nabla f(\bar{x}^*)\|}$ and define $\rho(t) = f(\bar{x}^* + t\bar{u})$. We have $\rho'(0) = \nabla f(\bar{x}^*)^T \bar{u} = - \parallel \nabla f(\bar{x}^*) \parallel < 0$.

By definition, $\rho'(0) = \lim_{t \to 0} \frac{\rho(t) - \rho(0)}{t}$, thus for all sufficiently small $t > 0$ we have that $\frac{\rho(t)-\rho(0)}{t} \leq \frac{\rho'(0)}{2} < 0$.

Therefore, $f(\bar{x}^* + t\bar{u}) - f(\bar{x}) < 0$. Therefore, $\bar{x}^*$ is not a minimizer, which contradicts our assumption that $\nabla f(\bar{x}^*) \neq 0$. ■

Note that this condition is not sufficient. For example, consider the following functions, $f(x) = x^3$ and $f(x) = -x^2$. For each of these functions, zero is a critical point, i.e. $\nabla f(0) = 0$, and yet it is not a minimizer.

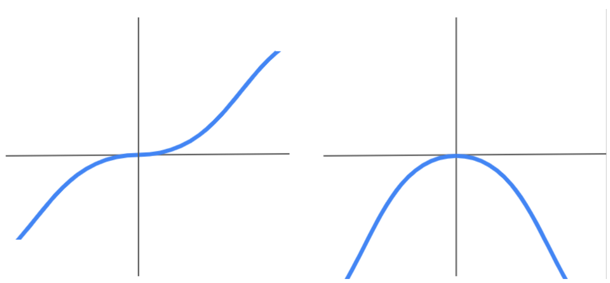

Figure 2.6: Examples of non-optimality that satisfy the 1st order necessary condition.

2.2.3 First order sufficient optimality condition for convex functions

We say that a function f is convex if $\forall x, y \in \mathbb{R}^d$ and for all$t \in [0, 1]$ we have:

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

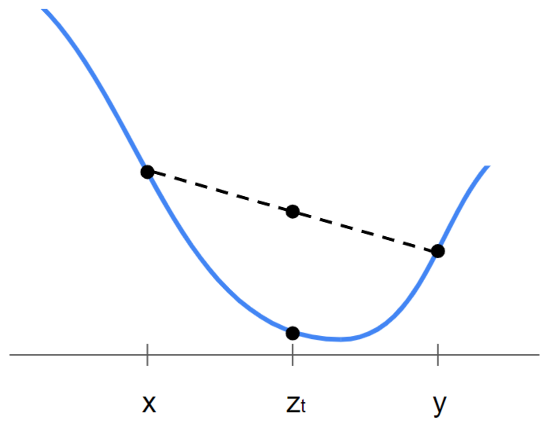

Figure 2.7: Visualization of convexity, where $z_t = tx + (1-t)y$, the equation left-hand side is a point on the function and the right-hand side is on the dotted line.

Theorem 2.5 *If f is differentiable and convex, then $\bar{x}^*$ is a global minimizer* $\iff \nabla f(\bar{x}^*) = 0$.