EECS 553: Machine Learning (ECE) **University of Michigan**

Unconstrained Optimization

Winter 2023 **Clayton Scott**

An *unconstrained optimization problem* has the form

$$\min_{\boldsymbol{u} \in \mathbb{R}^d} f(\boldsymbol{u})$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is called the *objective function.* A point $\boldsymbol{u}^* \in \mathbb{R}^d$ is called a *local minimizer* if $\exists r > 0$ such that $f(\boldsymbol{u}^*) \leq f(\boldsymbol{u})$ $\forall \boldsymbol{u}$ satisfying $\|\boldsymbol{u} - \boldsymbol{u}^*\| < r$. $\boldsymbol{u}^*$ is called a *global minimizer* if $f(\boldsymbol{u}^*) \leq f(\boldsymbol{u})$ $\forall \boldsymbol{u} \in \mathbb{R}^d$.

An understanding of basic optimization theory is essential for machine learning. Most of the algorithms we study will be formulated as solutions to optimization problems where the objective function measures the quality of the structure being learned.

1 First and Second Order Necessary Conditions

Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *gradient* and *Hessian* of f at $\boldsymbol{u} = [u_1 \ \cdots \ u_d]^T \in \mathbb{R}^d$ are defined by

$$\nabla f(\boldsymbol{u}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{u})}{\partial u_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{u})}{\partial u_d} \end{bmatrix}$$

and

$$\nabla^2 f(\boldsymbol{u}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{u})}{\partial u_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{u})}{\partial u_1 \partial u_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{u})}{\partial u_d \partial u_1} & \cdots & \frac{\partial^2 f(\boldsymbol{u})}{\partial u_d^2} \end{bmatrix}$$

provided the required partial derivatives exist. We say f is *differentiable* if $\nabla f(\boldsymbol{u})$ exists $\forall \boldsymbol{u} \in \mathbb{R}^d$, and *twice differentiable* if $\nabla^2 f(\boldsymbol{u})$ exists $\forall \boldsymbol{u} \in \mathbb{R}^d$. We say f is *twice continuously differentiable* if it is twice differentiable and all of the second derivatives are continuous. If f is twice continuously differentiable, then $\nabla^2 f(\boldsymbol{u})$ is symmetric $\forall \boldsymbol{u}$, i.e.,

$$\frac{\partial^2 f(\boldsymbol{u})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\boldsymbol{u})}{\partial x_j \partial x_i}$$
$$\forall \boldsymbol{u} \in \mathbb{R}^d$$
$$\forall i, j = 1, \ldots, d.$$

Property 1. If f is differentiable and $\boldsymbol{u}^*$ is a local minimizer of f, then $\nabla f(\boldsymbol{u}^*) = \mathbf{0}$.

This is called a *first-order necessary condition* for $\boldsymbol{u}^*$ to be a local minimizer. See Figure 1.

Proof. Define the function $\phi(t) = f(\boldsymbol{u}^* + \boldsymbol{v}t)$, where $\boldsymbol{v} \in \mathbb{R}^d$ is arbitrary. Since $\boldsymbol{u}^*$ is a local minimizer, we know that for $t > 0$ sufficiently small,

$$f(\boldsymbol{u}^* + \boldsymbol{v}t) \geq f(\boldsymbol{u}^*),$$

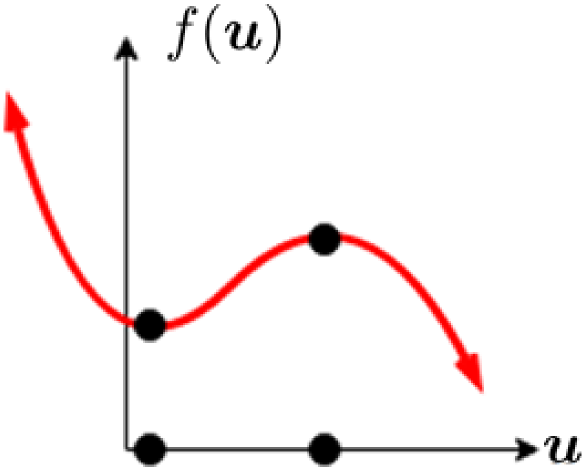

Figure 1: $\nabla f(\boldsymbol{u}^*) = \boldsymbol{0}$ is necessary but not sufficient for $\boldsymbol{u}^*$ to be a local minimizer.

and therefore,

$$\begin{aligned}\phi'(0) &= \lim_{t \searrow 0} \frac{f(\boldsymbol{u}^* + \boldsymbol{v}t) - f(\boldsymbol{u}^*)}{t} \\ &\geq 0.\end{aligned}$$

However, from the multivariable chain rule, we also know that

$$\phi'(0) = \langle \nabla f(\boldsymbol{u}^*), \boldsymbol{v} \rangle.$$

Therefore, $\langle \nabla f(\boldsymbol{u}^*), \boldsymbol{v} \rangle \geq 0$. Now choose $\boldsymbol{v} = -\nabla f(\boldsymbol{u}^*)$. Then

$$0 \leq \langle \nabla f(\boldsymbol{u}^*), -\nabla f(\boldsymbol{u}^*) \rangle = -\|\nabla f(\boldsymbol{u})\|^2 \leq 0,$$

so we must have $\nabla f(\boldsymbol{u}^*) = \boldsymbol{0}$. □

Property 2. If f is twice continuously differentiable and $\boldsymbol{u}^*$ is a local min, then $\nabla^2 f(\boldsymbol{u}^*)$ is positive semidefinite, i.e., $\boldsymbol{z}^T \nabla^2 f(\boldsymbol{u}^*) \boldsymbol{z} \geq 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^d$.

The proof is left as an exercise. This result generalizes the result from single-variable calculus that the second derivative is nonnegative at a local min. See Figure 2.

2 Convexity

We say that f is *convex* if

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v})$$

$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $t \in [0,1]$. We say f is *strictly convex* if

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) < tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v})$$

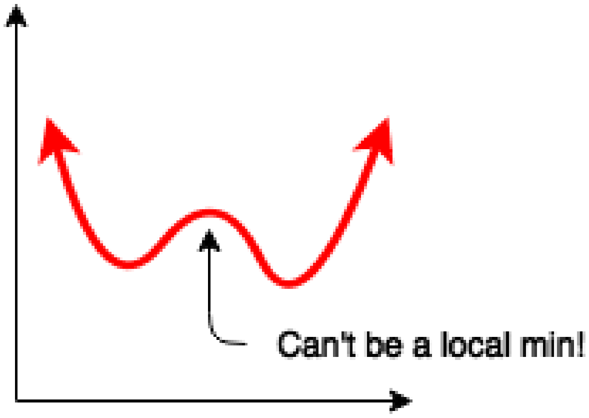

Figure 2: Illustration of second-order necessary condition.

$\forall \boldsymbol{u} \neq \boldsymbol{v}$ and $t \in (0,1)$. Geometrically, convexity means that for any two points on the graph of the function, the line segment connecting them is never below the graph. Strict convexity means that for any two distinct points on the graph of the function, the line segment connecting them is above the graph (except at its endpoints). See Figure 3.

If f is convex, the problem of minimizing f becomes easier to understand. Let's look at some basic properties.

Property 3. If f is convex, then every local min is a global min.

Proof. Suppose $\boldsymbol{u}^*$ is a local min but not a global min. Then $\exists \boldsymbol{v}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{v}^*) < f(\boldsymbol{u}^*)$. By convexity, $\forall t \in [0,1)$ we have

$$\begin{aligned} f(t\boldsymbol{u}^* + (1-t)\boldsymbol{v}^*) &\leq tf(\boldsymbol{u}^*) + (1-t)f(\boldsymbol{v}^*) \\ &< tf(\boldsymbol{u}^*) + (1-t)f(\boldsymbol{u}^*) \\ &= f(\boldsymbol{u}^*) \end{aligned}$$

Letting t approach 1, the point $t\boldsymbol{u}^* + (1-t)\boldsymbol{v}^*$ becomes arbitrarily close to $\boldsymbol{u}^*$. The above strict inequality contradicts local minimality of $\boldsymbol{u}^*$. Thus $\boldsymbol{u}^*$ is a global min. □

Property 4. If f is strictly convex, then f has at most one global min.

The proof is left as an exercise.

The following is a first-order characterization of convexity.

Property 5. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then f is convex iff $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$,

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \nabla f(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle.$$

Similarly, f is strictly convex iff $\forall \boldsymbol{u} \neq \boldsymbol{v}$,

$$f(\boldsymbol{v}) > f(\boldsymbol{u}) + \langle \nabla f(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle.$$

In words, this says that the tangent to the graph is always below the graph. See Figure 4.

Proof. First, assume f is convex. For any $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, t \in [0,1]$,

$$\begin{aligned} f(t\boldsymbol{v} + (1-t)\boldsymbol{u}) &\leq tf(\boldsymbol{v}) + (1-t)f(\boldsymbol{u}) \\ &= f(\boldsymbol{u}) + t(f(\boldsymbol{v}) - f(\boldsymbol{u})). \end{aligned} \tag{1}$$

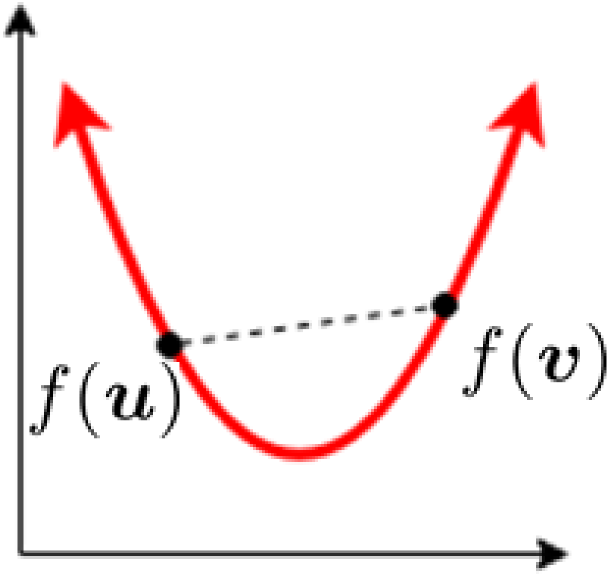

Figure 3: For a convex function, the line segment connecting two points on the graph is never below the graph.

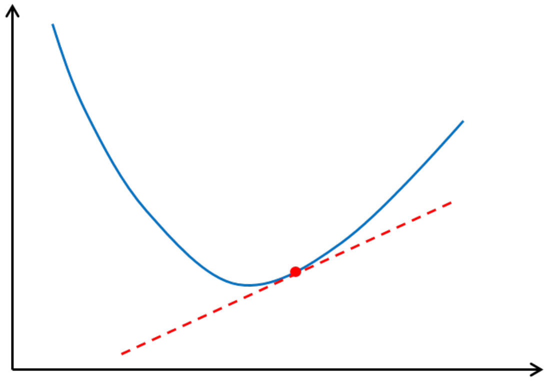

Figure 4: The graph is above the tangent line.

Rearranging,

$$\frac{f(\boldsymbol{u}+t(\boldsymbol{v}-\boldsymbol{u}))-f(\boldsymbol{u})}{t} \leq f(\boldsymbol{v})-f(\boldsymbol{u}).$$

The limit of the left-hand side (as $t \searrow 0$) is equal to $\langle \nabla f(\boldsymbol{u}), \boldsymbol{v}-\boldsymbol{u}\rangle$ by the multivariable chain rule. Therefore $f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \nabla f(\boldsymbol{u}), \boldsymbol{v}-\boldsymbol{u}\rangle$. Now suppose that $\forall \boldsymbol{u}, \boldsymbol{v}$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \nabla f(\boldsymbol{u}), \boldsymbol{v}-\boldsymbol{u}\rangle. \tag{2}$$

Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $t \in [0,1]$. Denote $\boldsymbol{z} = t\boldsymbol{u} + (1-t)\boldsymbol{v}$. Applying (2) twice we have

$$f(\boldsymbol{u}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{u}-\boldsymbol{z}\rangle \tag{3}$$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{v}-\boldsymbol{z}\rangle \tag{4}$$

Now multiply (3) by t, (4) by $(1-t)$, and add:

$$\begin{aligned} tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v}) &\geq f(\boldsymbol{z}) + < \nabla f(\boldsymbol{z}), t\boldsymbol{u} + (1-t)\boldsymbol{v} - \boldsymbol{z} > \\ &= f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \end{aligned}$$

as $t\boldsymbol{u} + (1-t)\boldsymbol{v} - \boldsymbol{z} = 0$. This establishes convexity. The proof of the second statement (strict convexity) is similar. □

For convex and differentiable f, the first order necessary condition is also sufficient

Property 6. Let f be convex and differentiable. Then $\boldsymbol{u}^*$ is a global min iff $\nabla f(\boldsymbol{u}^*) = 0$.

Proof. The forward implication follows from Property 1. The reverse implication follows immediately from Property 5. □

Property 7. Let f be twice continuously differentiable. Then

1. f is convex iff $\nabla^2 f(\boldsymbol{u})$ is positive semi-definite $\forall \boldsymbol{u} \in \mathbb{R}^d$.
2. f is strictly convex if $\nabla^2 f(\boldsymbol{u})$ is positive definite $\forall \boldsymbol{u} \in \mathbb{R}^d$.

The proof is left as an exercise.

Exercises

1. (☆☆☆) Prove Property 2. *Hint:* A multivariable version of Taylor's Theorem states that a twice continuously differentiable function can be expressed

$$f(\boldsymbol{u}) = f(\boldsymbol{v}) + \big\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \big\rangle + \frac{1}{2}\big\langle \boldsymbol{u} - \boldsymbol{v}, \nabla^2 f(\boldsymbol{v})(\boldsymbol{u} - \boldsymbol{v}) \big\rangle + o\left(\|\boldsymbol{u} - \boldsymbol{v}\|^2\right),$$

where $o(t)$ denotes a function satisfying $\lim\limits_{t \to 0} \dfrac{o(t)}{t} = 0$.

2. (★) Give an example of a function f such that $\nabla^2 f(\boldsymbol{u}^*)$ is positive semi-definite at some point $\boldsymbol{u}^*$, but $\boldsymbol{u}^*$ is not a local minimizer.

3. (★) Give an example of a function that is convex but not strictly convex.

4. (★★) Prove Property 4.

5. (★) Give an example of an f that is

   (a) convex and has more than one global min.

   (b) strictly convex and has no global min.

6. (★) Give an example of an f that is strictly convex such that $\nabla^2 f(\boldsymbol{u})$ is not positive definite for all $\boldsymbol{u}$.

7. (☆☆☆) Prove Property 7. *Hint:* Another multivariable version of Taylor's theorem states that a twice continuously differentiable function can be expressed

$$f(\boldsymbol{u}) = f(\boldsymbol{v}) + \big\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \big\rangle + \frac{1}{2}\left\langle \boldsymbol{u} - \boldsymbol{v}, \nabla^2 f\big(\boldsymbol{v} + t(\boldsymbol{u} - \boldsymbol{v})\big)(\boldsymbol{u} - \boldsymbol{v}) \right\rangle,$$

for some $t \in (0,1)$ possibly depending on $\boldsymbol{u}$ and $\boldsymbol{v}$.

8. (★) Prove that the sum of two convex functions is convex in two different ways.

   (a) First, use the definition of convexity.

   (b) Second, use the Hessian. You may assume that the functions are twice continuously differentiable.

9. (★★) Prove the following statements, or disprove them by providing a counterexample.

   (a) The product of two convex functions is convex.

   (b) The composition of two convex functions is convex.

10. (★★) Consider the function $f(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u} + \boldsymbol{b}^T \boldsymbol{u} + c$, where $\boldsymbol{A}$ is a symmetric $d \times d$ matrix. Derive the Hessian of f. Under what conditions on $\boldsymbol{A}$ is f convex? Strictly convex?