Lecture 10

Optimization, Part 1

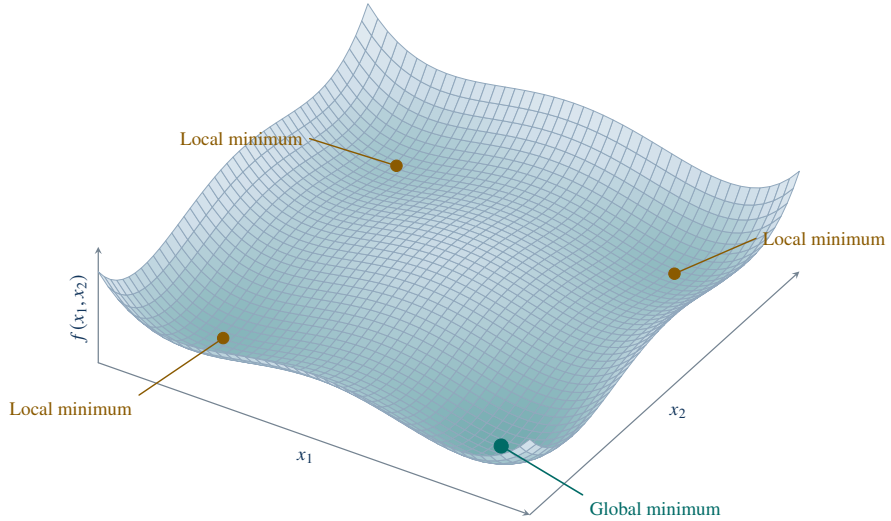
Virginia Tech Department of Computer Science
Fall 2026

Unconstrained optimization problems

The variables are given by a vector $\boldsymbol{x} \in \mathbb{R}^n$. The **objective function** is $f : \mathbb{R}^n \to \mathbb{R}$.
The problem is to minimize the function f over all $\boldsymbol{x} \in \mathbb{R}$:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

"Unconstrained" means that no additional equations or inequalities restrict the variables.



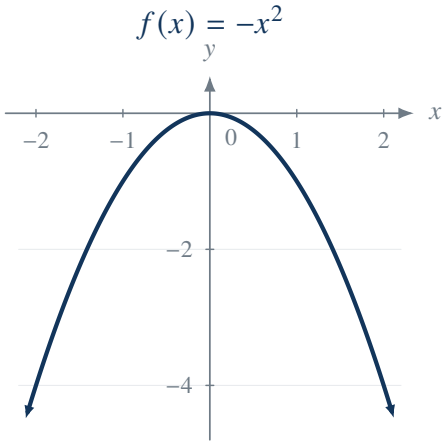
There are many algorithms for solving optimization problems. Today we will focus on **gradient descent**.

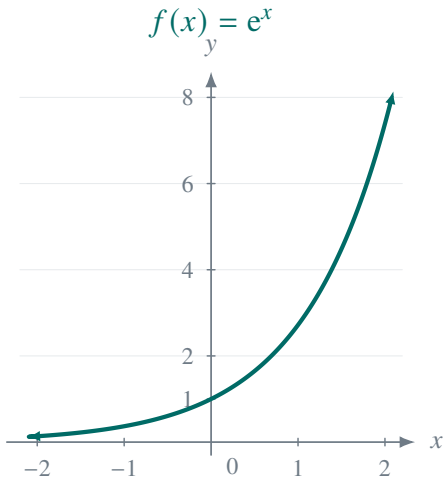
Local and global minimizers

A **global minimizer** $\boldsymbol{x}_*$ satisfies $f(\boldsymbol{x}_*) \leq f(\boldsymbol{x})$ *for every* $\boldsymbol{x} \in \mathbb{R}^n$.

A **local minimizer** satisfies the same inequality whenever $\|\boldsymbol{x} - \boldsymbol{x}_*\|_2 < r$ for some $r > 0$.
A local minimum is the best among nearby inputs.

A local minimum can be worse than a distant minimum. The function $f(x) = -x^2$ has no lower bound. The function $f(x) = \mathrm{e}^x$ has approaches arbitrarily close to 0 but never attains it.





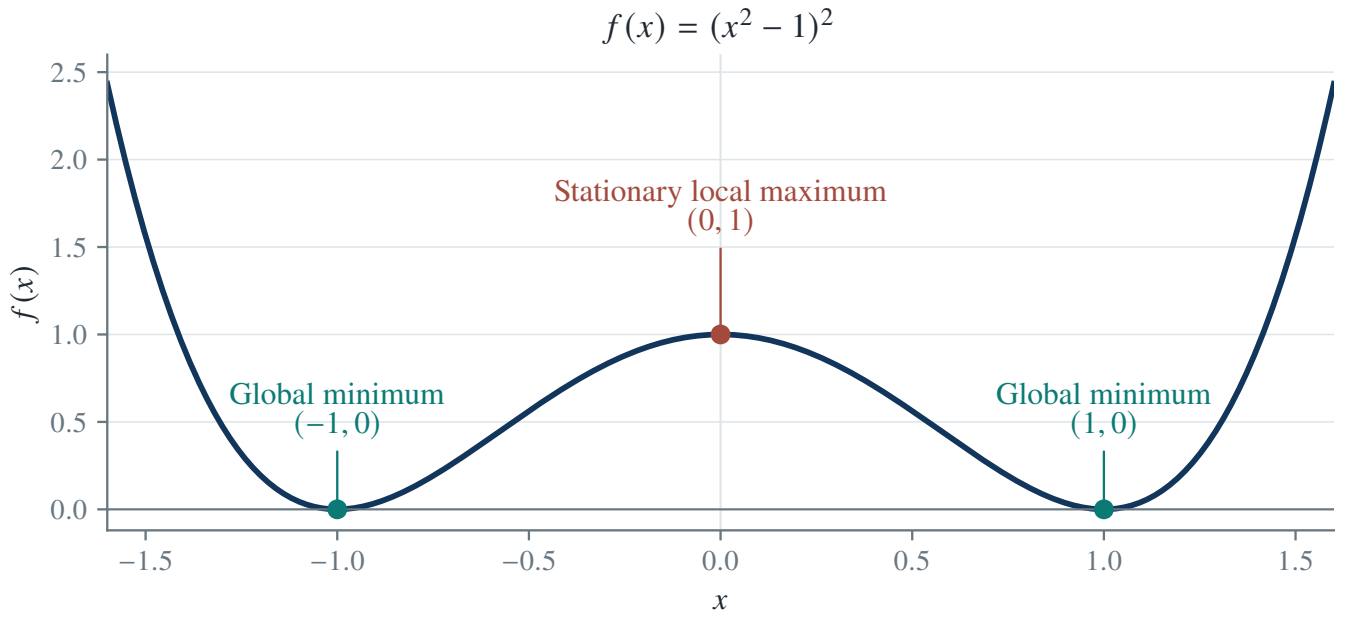
Single-variable optimality tests

Let f be twice continuously differentiable near x_*. A **stationary point** satisfies $f'(x_*) = 0$.

$$\begin{aligned} \text{Local minimum} &\implies f'(x_*) = 0, \quad f''(x_*) \geq 0. \\ f'(x_*) = 0,\ f''(x_*) > 0 &\implies \text{strict local minimum}. \\ f'(x_*) = 0,\ f''(x_*) < 0 &\implies \text{strict local maximum}. \end{aligned}$$

When $f''(x_*) = 0$, the second-derivative test is inconclusive. At 0, the functions x^4, $-x^4$, and x^3 have vanishing first and second derivatives but give a minimum, a maximum, and neither.

Example: stationary points



For $f(x) = (x^2 - 1)^2$, all three marked points are stationary. Only $x = \pm 1$ minimize the objective.

Example: stationary points

$$f(x) = (x^2 - 1)^2,$$
$$f'(x) = 4x(x^2 - 1), \qquad f''(x) = 12x^2 - 4.$$

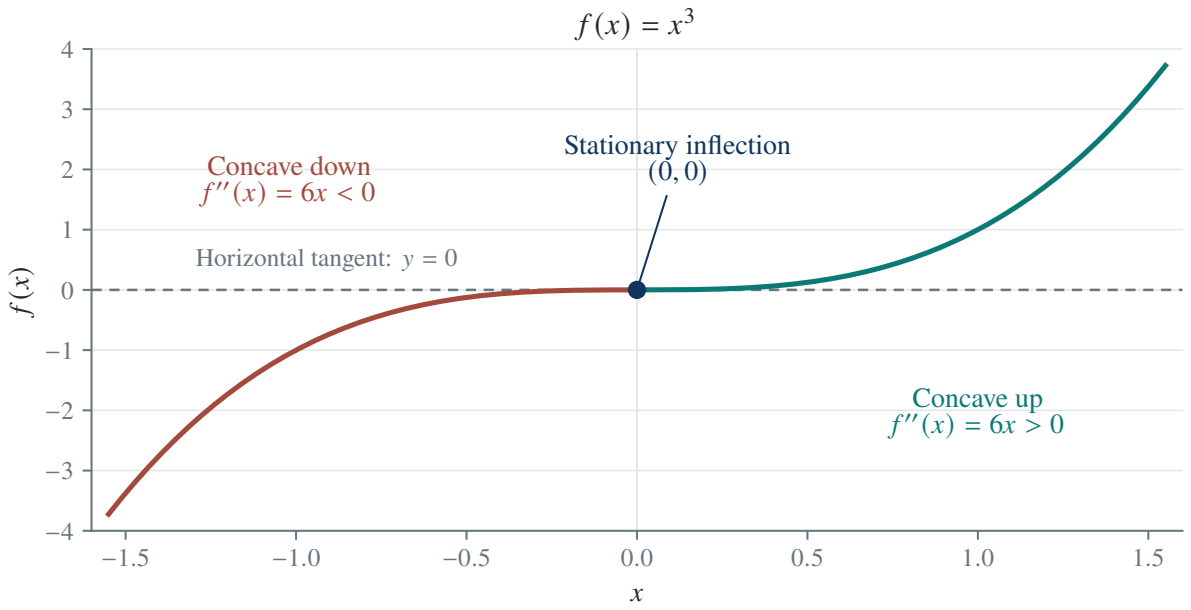
Solving $f'(x) = 0$ gives $x = -1, 0, 1$. Since $f''(\pm 1) = 8 > 0$, both -1 and 1 are strict local minima.

They are also global minima: $f(x) \geq 0$ everywhere and $f(\pm 1) = 0$.

At 0, $f''(0) = -4 < 0$, so the point is a strict local maximum. There is no global maximum because $f(x) \to +\infty$ as $|x| \to \infty$.

Example: a stationary inflection point

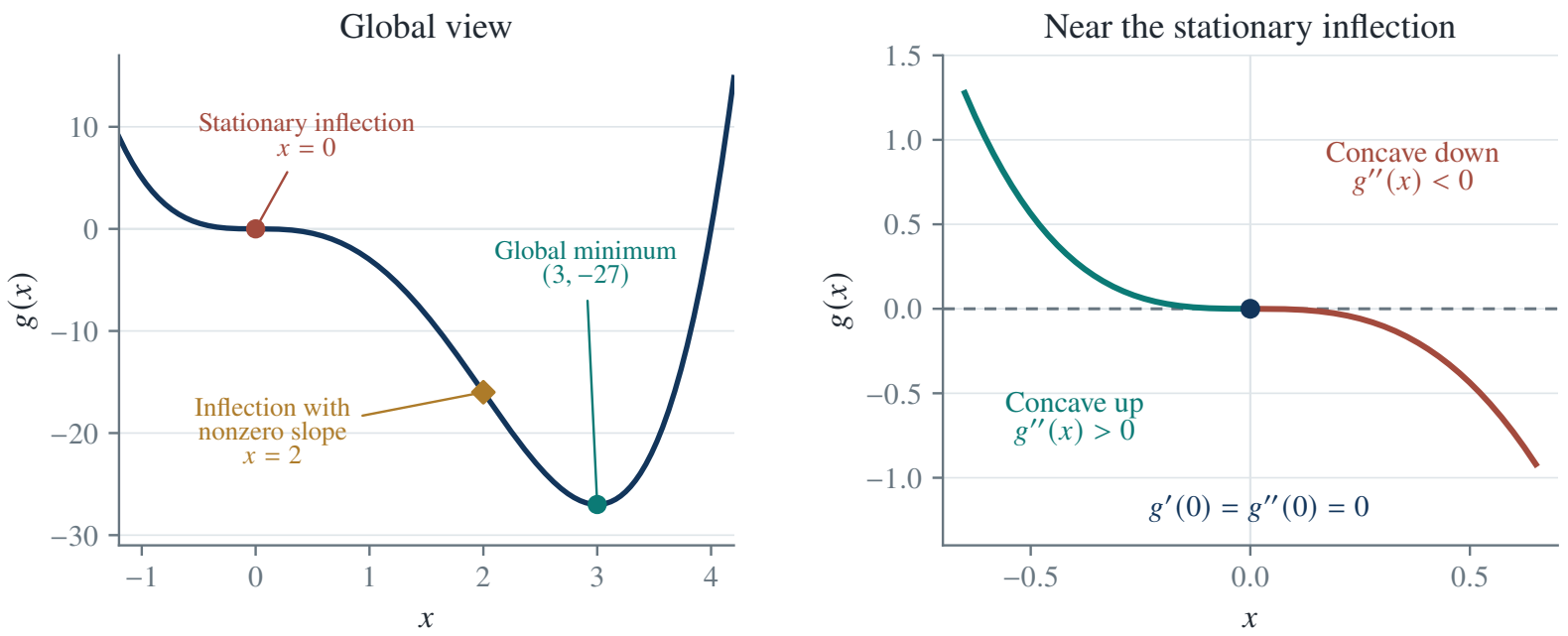
$$f(x) = x^3, \qquad f'(x) = 3x^2, \qquad f''(x) = 6x, \qquad f'''(x) = 6.$$



At 0, both the first and second derivatives vanish. For $h > 0, f(-h) < f(0) < f(h)$, so there is neither a minimum nor a maximum. Concavity changes from down to up: this is a stationary inflection point.

Example: a function with all three behaviors

$$g(x) = x^4 - 4x^3$$



$g(x) = x^4 - 4x^3$: stationary inflection at 0, non-stationary inflection at 2, and unique global minimum at 3.

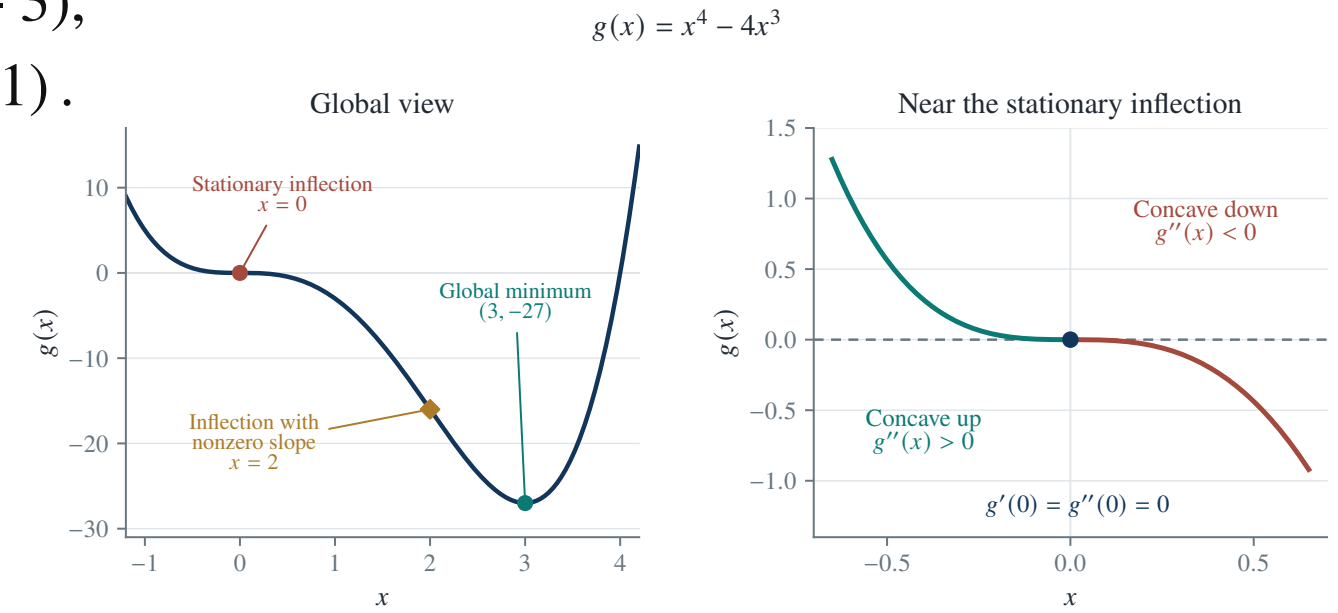
Example: a function with all three behaviors

$$g(x) = x^4 - 4x^3, \qquad g'(x) = 4x^2(x-3),$$
$$g''(x) = 12x(x-2), \qquad g'''(x) = 24(x-1).$$

The stationary points are 0 and 3. At 0, $g'(0) = g''(0) = 0$, while $g'''(0) = -24$.

Zero is neither a minimum nor a maximum.

Since g'' changes from positive to negative there, it is a stationary inflection point.



Example: a function with all three behaviors

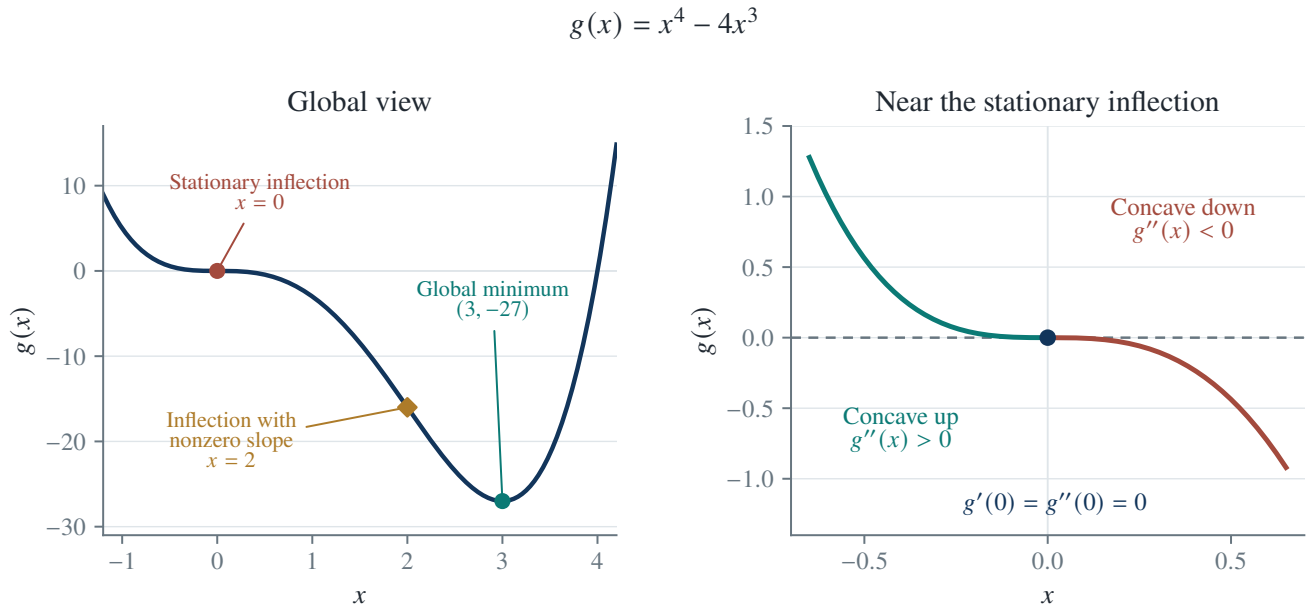
At $x = 2$, the second derivative changes from negative to positive. This is another inflection point, but it is not stationary:

$$g(2) = -16, \qquad g'(2) = -16 \neq 0.$$

At $x = 3$, $g''(3) = 36 > 0$, so there is a strict local minimum. The identity

$$g(x) + 27 = (x-3)^2\left[(x+1)^2 + 2\right] \geq 0$$

shows that this minimum is uniquely global: $g(x) \geq -27$, with equality only at $x = 3$.



Newton's method for finding stationary points

$$\text{Solve } f(x) = 0: \quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

$$\text{Solve } f'(x) = 0: \quad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

The second update seeks a stationary point. It does not distinguish a minimum from a maximum.

Multi-variable derivatives: the gradient and Hessian

For $f : \mathbb{R}^n \to \mathbb{R}$, the **gradient vector** collects first derivatives and the **Hessian matrix** collects second derivatives.

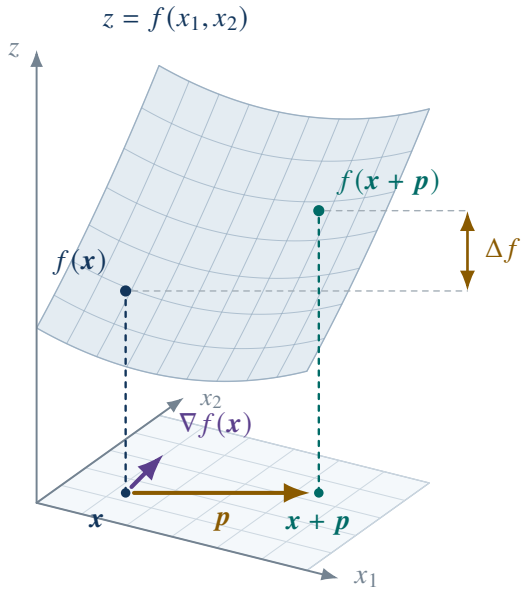
$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}, \qquad H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) \right)_{i,j=1}^{n} .$$

For twice continuously differentiable f, the Hessian is symmetric. The gradient describes local slope; the Hessian describes curvature in different directions.

$$f(\boldsymbol{x} + \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^{\mathsf{T}} H(\boldsymbol{x}) \boldsymbol{p} + o(\|\boldsymbol{p}\|_2^2) .$$

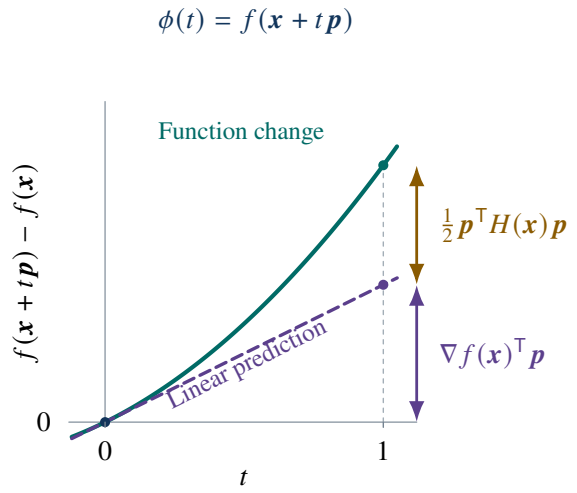
Taylor expansion at a point

$$f(\boldsymbol{x}+\boldsymbol{p}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^{\mathsf{T}}H(\boldsymbol{x})\boldsymbol{p} + o(\|\boldsymbol{p}\|_2^2)\,.$$



Gradient: direction of steepest increase

Hessian: curvature in each input direction

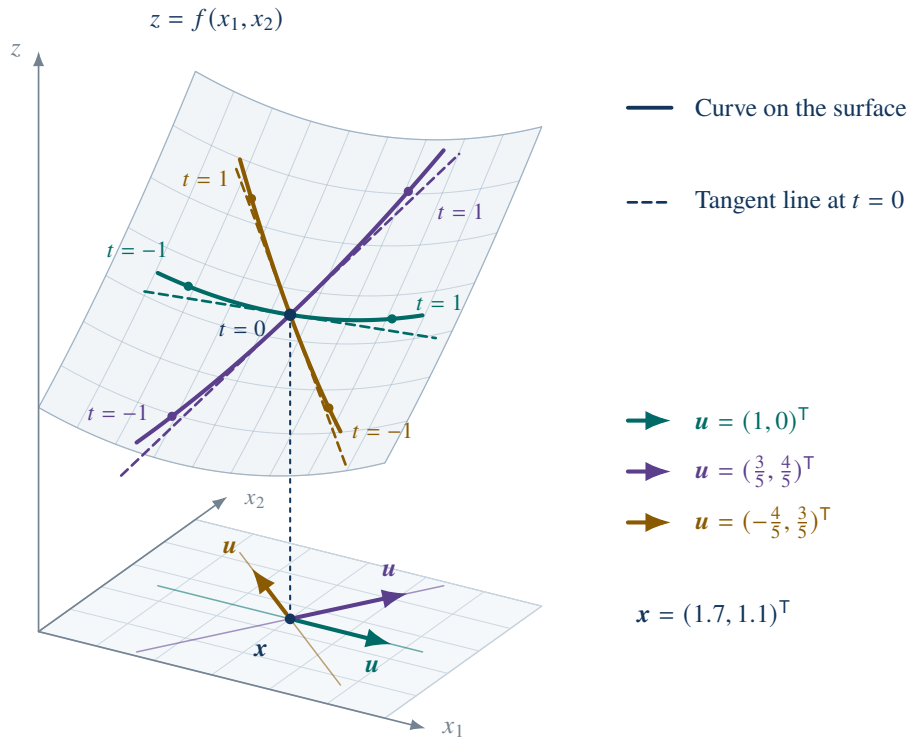


$\phi'(0) = \nabla f(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{p}$

$\phi''(0) = \boldsymbol{p}^{\mathsf{T}}H(\boldsymbol{x})\boldsymbol{p}$

Which direction to go?

Fix a point $\boldsymbol{x}$ and a unit direction $\boldsymbol{u}$. The line $\boldsymbol{x} + t\boldsymbol{u}$ passes through $\boldsymbol{x}$ at $t = 0$. The scalar t is a line parameter; the distance from $\boldsymbol{x}$ is $|t|$.



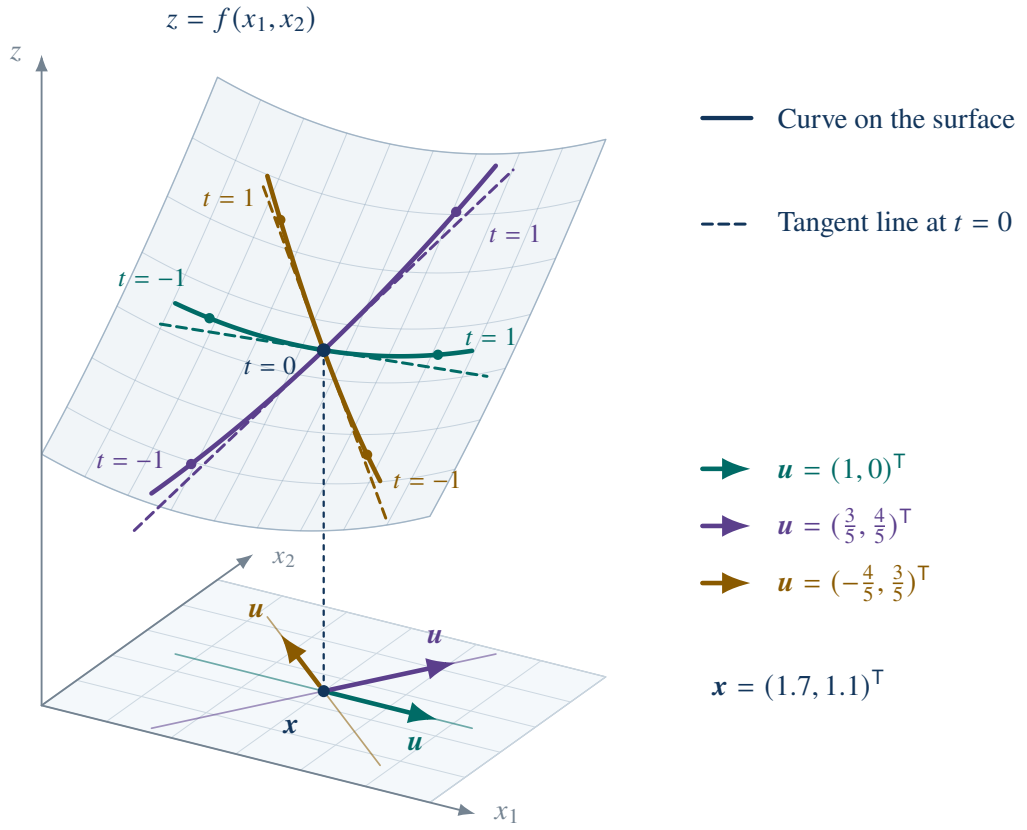
For optimization, we want to move in a direction that takes us towards the minimum. Which direction is that?

The gradient points in the steepest directions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. For a unit vector $\boldsymbol{u}$, the directional derivative is the rate of change in the direction of $\boldsymbol{u}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x} + t\boldsymbol{u})\bigg|_{t=0} = \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{u} \,.$$

If $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$, the greatest and smallest rates occur in the directions of the gradient.



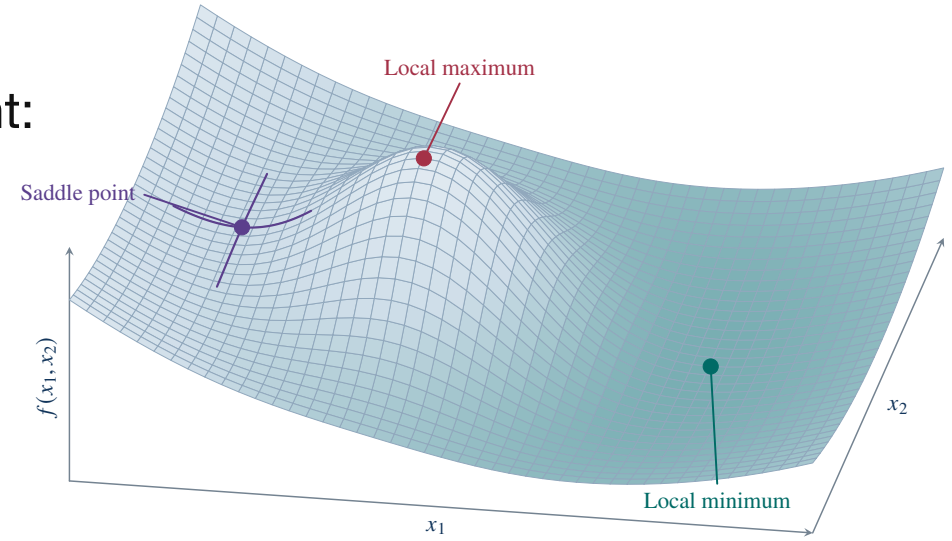
Classifying stationary points for multi-variable functions

A **stationary point** is given by $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$.

Suppose f is twice continuously differentiable near $\boldsymbol{x}_* \in \mathbb{R}^n$.

At a stationary point, the following Hessian tests are sufficient:

$$H(\boldsymbol{x}_*) \text{ positive definite} \implies \text{strict local minimum}.$$
$$H(\boldsymbol{x}_*) \text{ negative definite} \implies \text{strict local maximum}.$$
$$H(\boldsymbol{x}_*) \text{ indefinite} \implies \text{saddle point}.$$



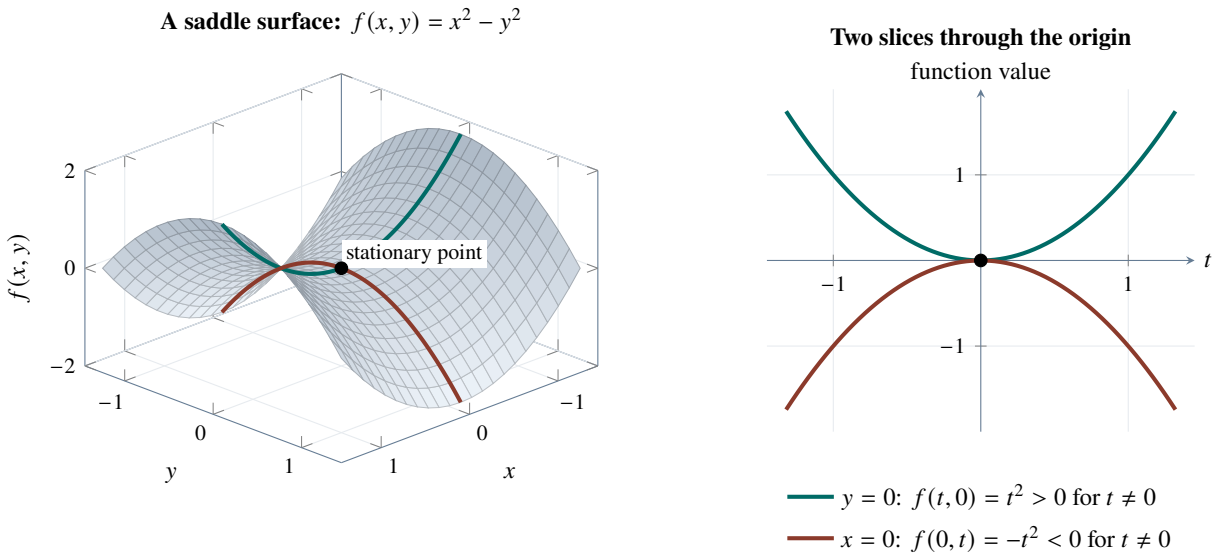
Positive definite means all eigenvalues are positive.

Negative definite has all negative eigenvalues.

Indefinite has both positive and negative eigenvalues.

A saddle has both higher and lower values nearby

$$f(x, y) = x^2 - y^2, \qquad f(t,0) = t^2, \qquad f(0,t) = -t^2 .$$

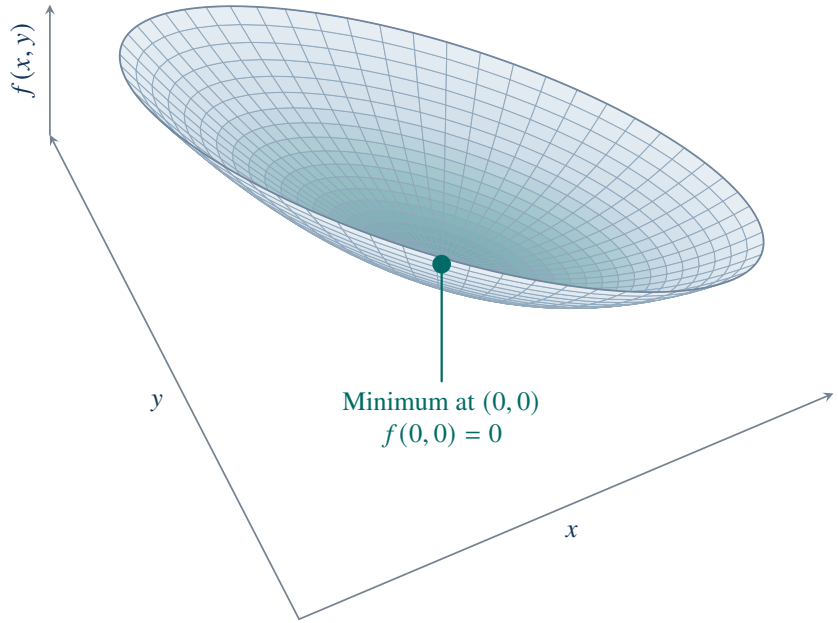


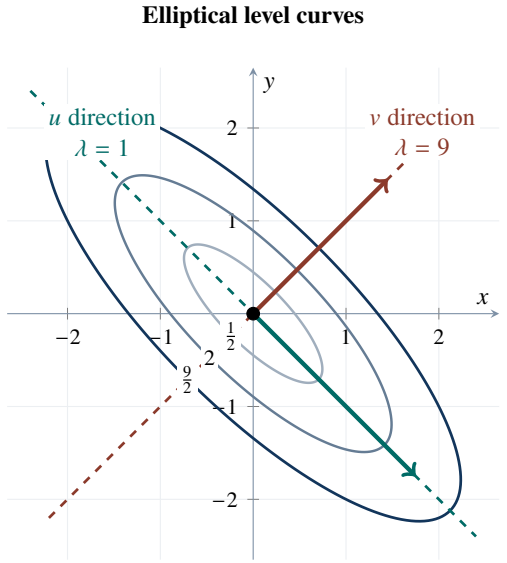
The origin is stationary, yet every neighborhood contains both higher and lower objective values. Its Hessian has eigenvalues 2 and -2.

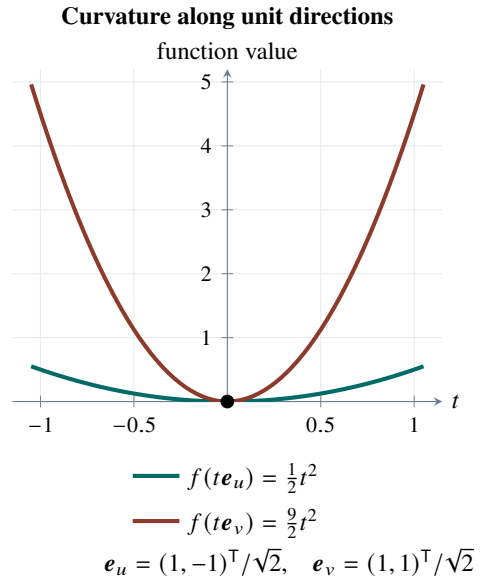
Example: stationary global minimum

$$f(x,y) = \frac{1}{2}(5x^2 + 8xy + 5y^2), \quad \nabla f(x,y) = \begin{pmatrix} 5x + 4y \\ 4x + 5y \end{pmatrix}, \quad H = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}.$$

The origin is the unique global minimizer.







Convexity and strict convexity

What are some *sufficient conditions* for local and/or global minima?

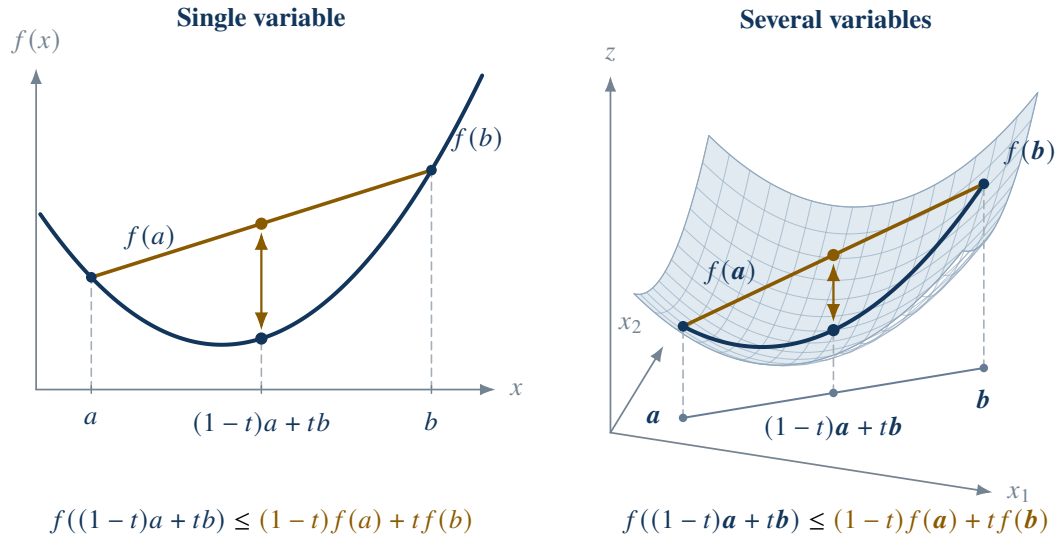
A function is **convex** if, for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $0 \leq t \leq 1$,

$$f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \leq (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y})\,.$$

The graph lies on or below every chord. **Strict convexity** makes this inequality strict for distinct points and $0 < t < 1$.

For a differentiable function, convexity is equivalent to every tangent plane being a global lower bound:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x})\,.$$



Every chord lies on or above the graph.

The inequality must hold for every pair of input points and every $t \in [0, 1]$.

What convexity guarantees

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Every local minimizer is a global minimizer.

If f is also differentiable, then

$$\boldsymbol{x}_* \textbf{ is a global minimizer} \quad \Longleftrightarrow \quad \nabla f(\boldsymbol{x}_*) = \mathbf{0}\,.$$

If f is **strictly convex**, it has at most one global minimizer. If a minimizer exists, it is unique.

Neither convexity nor strict convexity alone guarantees that a minimizer exists.

Hessian criteria for convexity

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable.

Convexity is equivalent to the Hessian being positive semidefinite at every point.

A positive definite Hessian at every point implies strict convexity. This condition is sufficient but not necessary.

Example: many global minimizers

$$f(x, y) = x^2, \qquad H(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

The Hessian is positive semidefinite everywhere, so the function is convex.

Since $f(x, y) \geq 0$ and $f(0,y) = 0$ for every y, every point on the line $x = 0$ is a global minimizer.

The function is constant along this line, so it is not strictly convex. The zero Hessian eigenvalue also rules out strong convexity.

Strong convexity provides existence and uniqueness

A differentiable function is μ-**strongly convex** if one fixed $\mu > 0$ satisfies, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Its graph stays above an upward-opening quadratic surface based at every point. The extra term prevents arbitrarily flat curvature.

For a fixed $\mu > 0$, μ-strong convexity is equivalent to every eigenvalue of every Hessian being at least equal to μ: $H(\boldsymbol{x}) - \mu\mathbb{1} \succeq 0$ *for every* $\boldsymbol{x} \in \mathbb{R}^n$.

Strong convexity implies strict convexity, which implies convexity. The converse implications do not generally hold.

Theorem: If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and μ-strongly convex for some $\mu > 0$, then it attains a unique global minimum at $\boldsymbol{x}_*$.

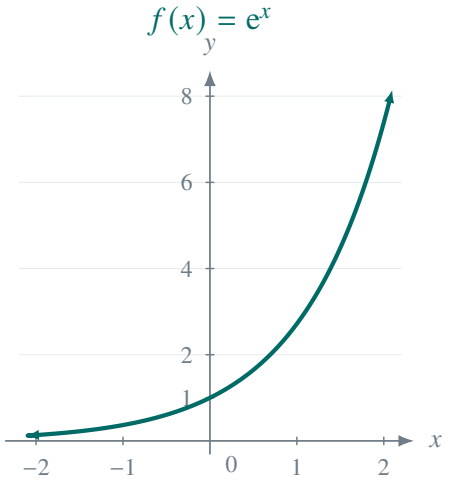
Strict convexity does not ensure existence

$$f(x) = \mathrm{e}^{x}, \qquad f''(x) = \mathrm{e}^{x} > 0 \quad \textit{for every } x \in \mathbb{R}\,.$$

The function is strictly convex, but $f(x) > 0$ at every finite input and $f(x) \to 0$ as $x \to -\infty$. Its infimum is 0 and no minimizer exists.

It is not strongly convex on $\mathbb{R}$: the positive second derivative approaches zero, so no fixed positive lower bound works everywhere.

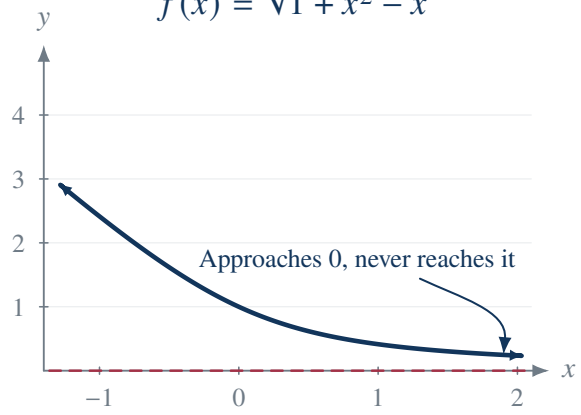
Strict convexity guarantees uniqueness only if a minimizer exists.



Example: a strictly convex function with no minimum

Strictly convex: no minimum

$$f(x) = \sqrt{1+x^2} - x$$

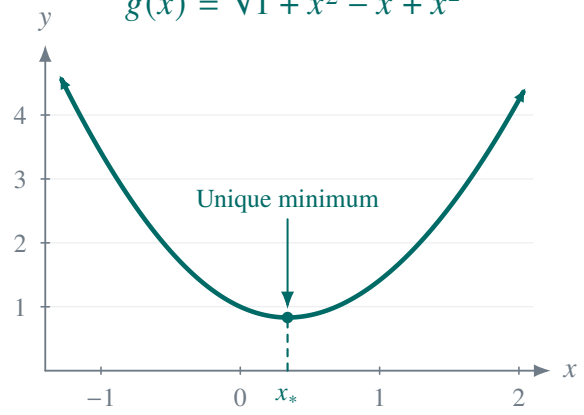


$$f''(x) = \frac{1}{(1+x^2)^{3/2}} > 0$$

Convexity alone does not guarantee existence.

Strongly convex: a unique minimum

$$g(x) = \sqrt{1+x^2} - x + x^2$$



$$g''(x) = \frac{1}{(1+x^2)^{3/2}} + 2 \geq 2$$

Strong convexity guarantees existence and uniqueness.

$$g(x) = \sqrt{1+x^2} - x + x^2, \qquad g''(x) = \frac{1}{(1+x^2)^{3/2}} + 2 \geq 2.$$

The function g is 2-strongly convex on $\mathbb{R}$. It attains a unique global minimum at approximately $x_* = 0.3393$.

Example: a strictly convex function with no minimum

$$f(x) = \sqrt{1 + x^2} - x, \qquad f''(x) = \frac{1}{(1 + x^2)^{3/2}} > 0.$$

The positive second derivative gives strict convexity. Rationalizing the expression yields

$$f(x) = \frac{1}{\sqrt{1 + x^2} + x} > 0.$$

As $x \to +\infty$, the denominator grows without bound and $f(x) \to 0$. Every finite candidate can be improved, so the function has no global minimum.

Using gradient descent to find a minimum point

How do we actually find a minimum point?

1. Choose $\boldsymbol{x}_0$, a gradient tolerance, and an iteration limit.

2. Evaluate $f(\boldsymbol{x}_k)$ and $\nabla f(\boldsymbol{x}_k)$. Stop if the gradient norm is below the tolerance.

3. Choose the **step size** $\alpha_k > 0$, update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$, and repeat.

A small gradient reports approximate stationarity. It does not distinguish a minimum, maximum, or saddle.

Several initial points may reveal different local minima.

How do we pick the step size?

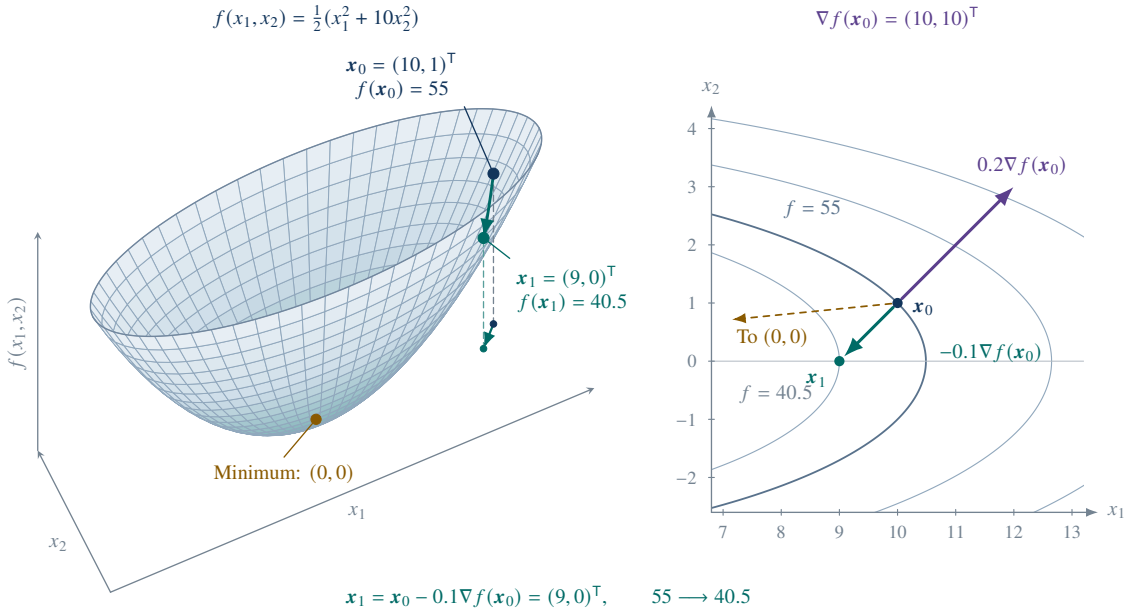
Example: One step of gradient descent

$$f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2), \qquad \boldsymbol{x}_0 = \begin{pmatrix} 10 \\ 1 \end{pmatrix}.$$

$$-\nabla f(\boldsymbol{x}_0) = \begin{pmatrix} -10 \\ -10 \end{pmatrix}, \qquad \mathbf{0} - \boldsymbol{x}_0 = \begin{pmatrix} -10 \\ -1 \end{pmatrix}.$$

$$\alpha_0 = 0.1 : \boldsymbol{x}_1 = \begin{pmatrix} 9 \\ 0 \end{pmatrix}$$

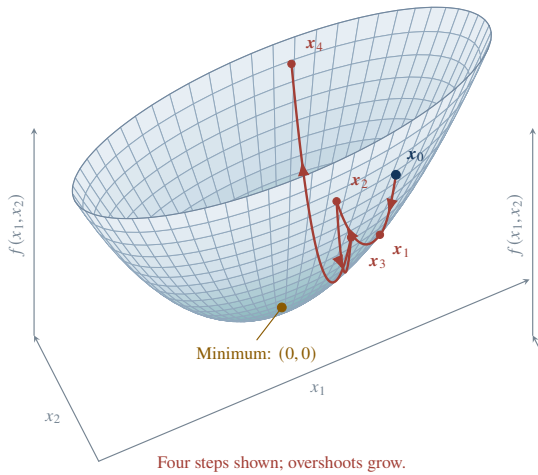
$$f(\boldsymbol{x}_0) = 55, f(\boldsymbol{x}_1) = 40.5.$$



The step size matters!

$f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2), \qquad \boldsymbol{x}_0 = (10, 1)^\mathsf{T}$

Large step: $\alpha = 0.25$

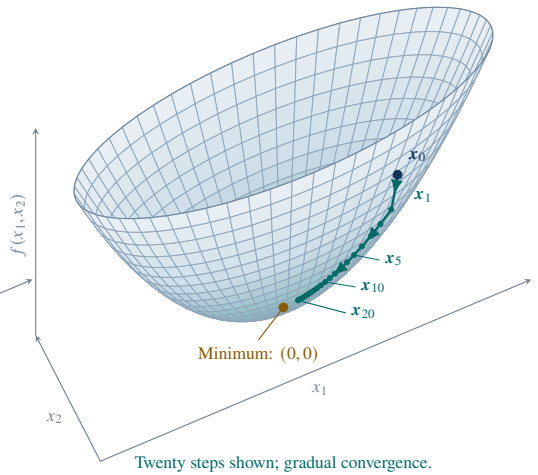


Four steps shown; overshoots grow.

$f(\boldsymbol{x}_k)$: $55 \to 39.38 \to 41.13 \to 65.85 \to 133.15$

$f(\boldsymbol{x}_k) \longrightarrow \infty$

Smaller step: $\alpha = 0.10$



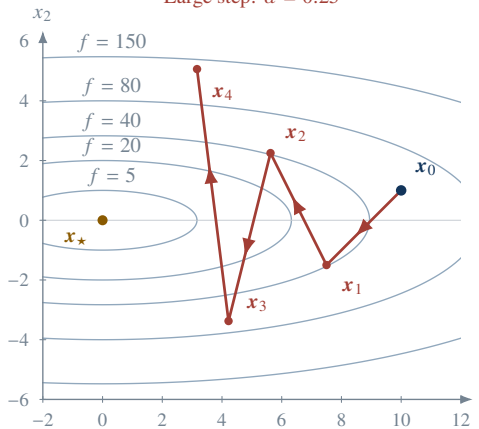
Twenty steps shown; gradual convergence.

$f(\boldsymbol{x}_{20}) \approx 0.739$

$\boldsymbol{x}_k \longrightarrow (0, 0)^\mathsf{T}$

$f(x_1, x_2) = \frac{1}{2}(x_1^2 + 10x_2^2), \qquad \boldsymbol{x}_0 = (10, 1)^\mathsf{T}$

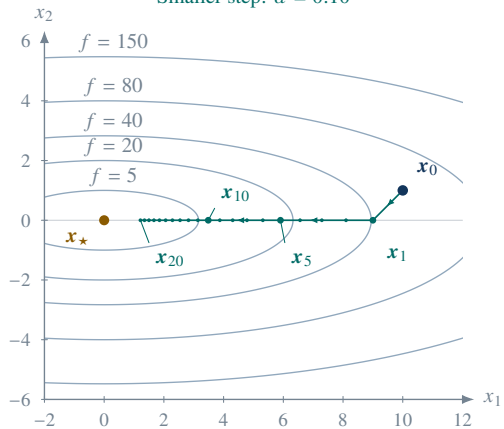
Large step: $\alpha = 0.25$



Four steps shown; overshoots grow.

$f(\boldsymbol{x}_4) \approx 133.15, \qquad f(\boldsymbol{x}_k) \longrightarrow \infty$

Smaller step: $\alpha = 0.10$



Twenty steps shown; gradual convergence.

$f(\boldsymbol{x}_{20}) \approx 0.739, \qquad \boldsymbol{x}_k \longrightarrow \boldsymbol{x}_\star$

Minimum: $\boldsymbol{x}_\star = (0, 0)^\mathsf{T}$

Example: constant steps for a scalar quadratic

Let $f(x) = \frac{a}{2}(x-b)^2$, where $a > 0$, and use a constant step $\alpha > 0$.

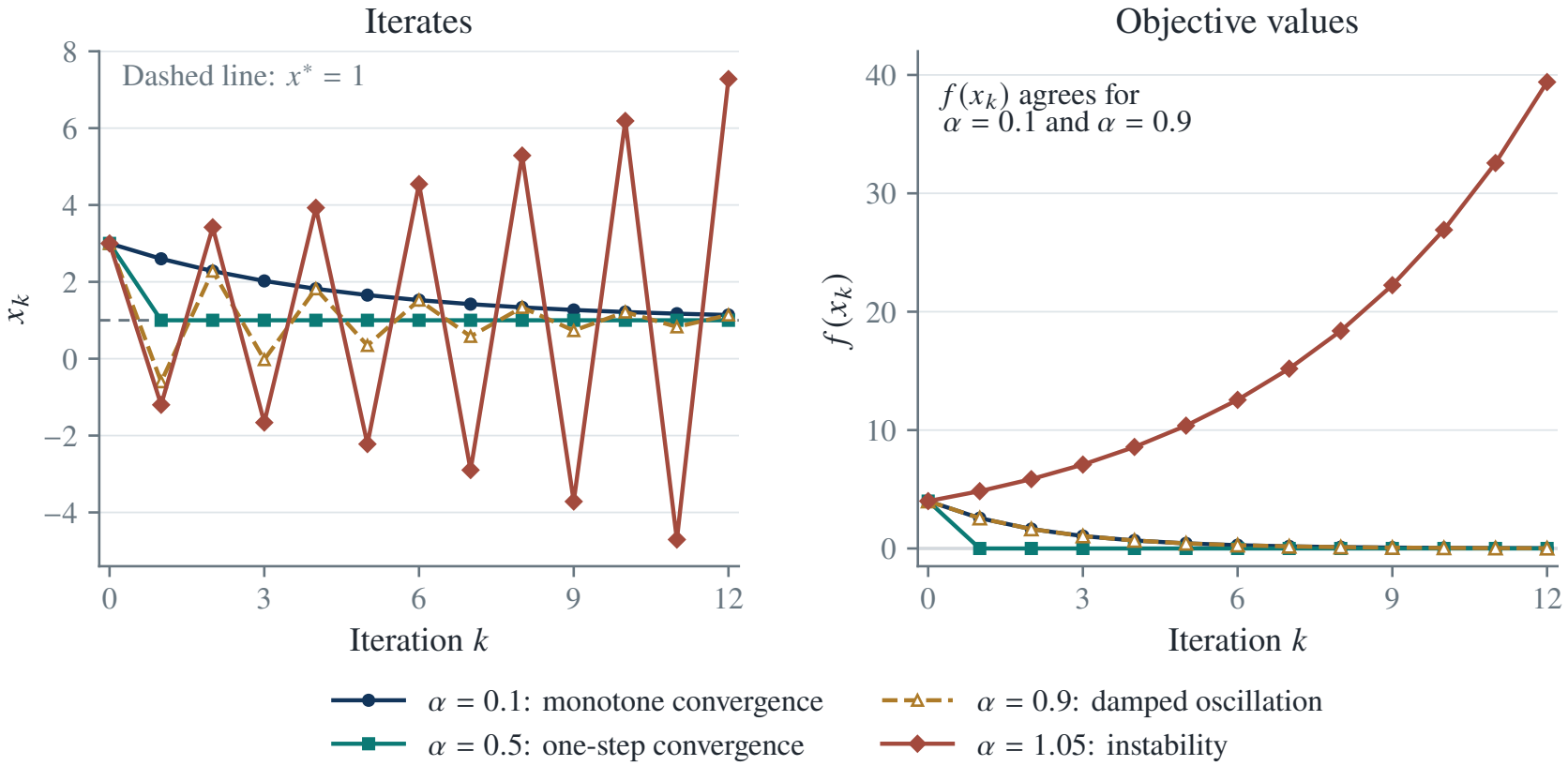
$$x_{k+1} - b = (1 - \alpha a)(x_k - b),$$
$$x_k - b = (1 - \alpha a)^k (x_0 - b), \qquad k \geq 1.$$

The iterates converge to b from every initial point if and only if

$$0 < \alpha < \frac{2}{a}.$$

The step size matters!

$f(x) = (x-1)^2$, $x_0 = 3$: here $a = 2$, so the stable range is $0 < \alpha < 1$.



Smoothness limits changes in the gradient

A differentiable function is L-smooth if its gradient is L**-Lipschitz continuous**, with $L > 0$:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2 \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n .$$

For a twice continuously differentiable function, $\|H(\boldsymbol{x})\|_2 \leq L$ everywhere is sufficient.

$$\|H\|_2 = \max_{\|\boldsymbol{p}\|_2=1} \|H\boldsymbol{p}\|_2.$$

Theorem: Descent lemma and a gradient step

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable with an L-Lipschitz gradient. For every point $\boldsymbol{x}$ and displacement $\boldsymbol{p}$,

$$f(\boldsymbol{x} + \boldsymbol{p}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{p} + \frac{L}{2} \|\boldsymbol{p}\|_2^2.$$

For every $\alpha > 0$,

$$f(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})) \leq f(\boldsymbol{x}) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(\boldsymbol{x})\|_2^2.$$

Thus $0 < \alpha < 2/L$ guarantees strict decrease whenever $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$.

A useful fixed step

Taking $\alpha = 1/L$ gives a simple quantitative decrease at every iteration:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k),$$

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

A large gradient forces a substantial decrease. A finite lower bound on the objective limits how long this can continue.

Theorem: Gradient descent for convex objective functions

Suppose f is convex, has an L-Lipschitz gradient on $\mathbb{R}^n$, and attains a global minimum at $\boldsymbol{x}_\star$.

Gradient descent with step size $1/L$ satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{2k}, \qquad k \geq 1.$$

Doubling the iteration count halves this upper bound. A particular run may converge faster.

Theorem: Gradient descent for strongly convex objective functions

Suppose f is μ-strongly convex and has an L-Lipschitz gradient on $\mathbb{R}^n$, with $\mu > 0$.

Then $\mu \leq L$, and f has a unique global minimizer $\boldsymbol{x}_\star$. With step size $1/L$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \leq \left(1 - \frac{\mu}{L}\right)^k \big(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_\star)\big), \qquad k \geq 1.$$

The points $\boldsymbol{x}_k$ also converge to $\boldsymbol{x}_\star$.