

L08

Strong Convexity and Gradient Descent

CS 169 / CS 268
Introduction to Optimization

Ioannis Panageas

Fall 2026

Recall: descent and convexity

For an L -smooth function, the step $\eta = 1/L$ gives:

$$f(x_{k+1}) \leq f(x_k) - \frac{[f'(x_k)]^2}{2L}.$$

If f is convex and has a minimizer x^* , then for $K \geq 1$:

$$f(x_K) - f(x^*) \leq \frac{L(x_0 - x^*)^2}{2K}.$$

Can additional curvature give a **faster rate**?

Strong convexity in one variable

A differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ is μ -strongly convex if $\mu > 0$ and, for every x, y ,

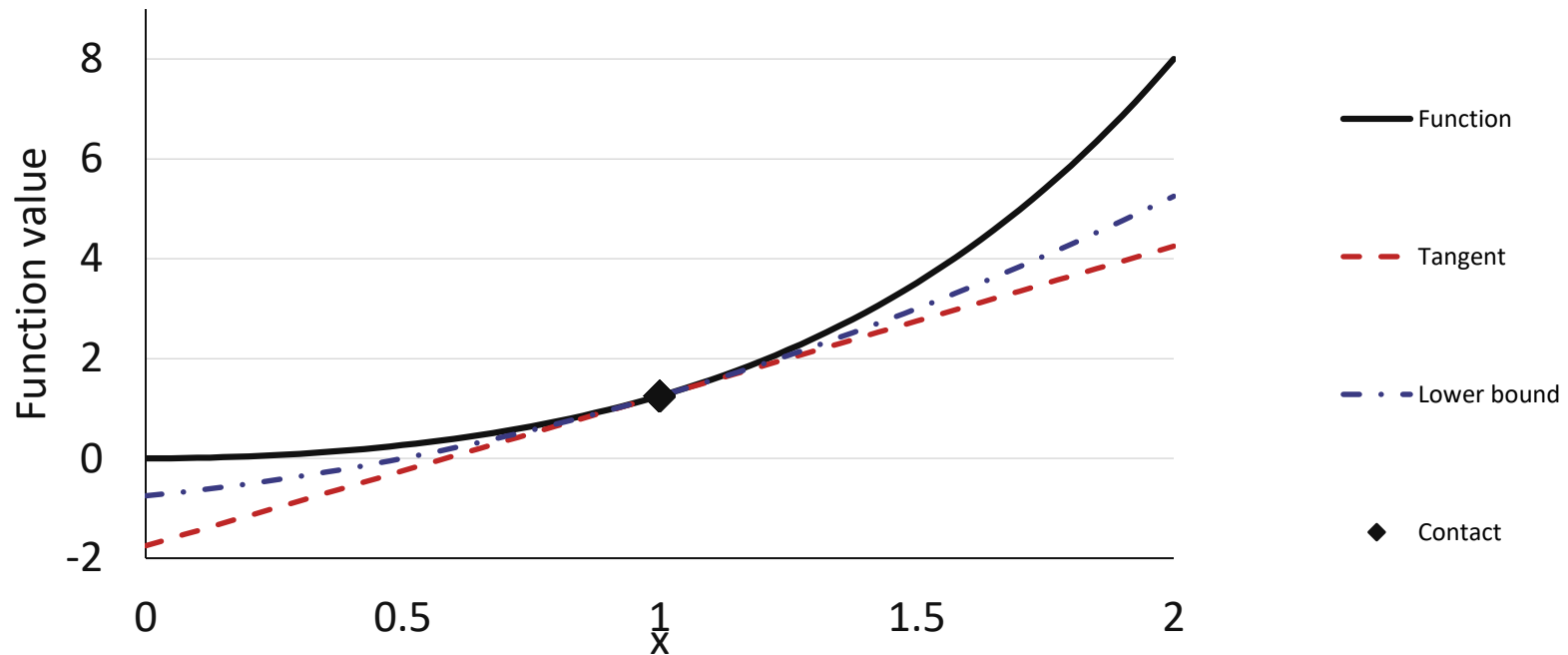
$$f(y) \geq f(x) + f'(x)(y - x) + \frac{\mu}{2}(y - x)^2.$$

The graph lies above a **quadratic lower bound**. This strengthens the tangent-line lower bound.

The same positive μ must work for every pair x, y .

A quadratic lower bound

$$f(x) = x^2 + \frac{x^4}{4}, \quad x_0 = 1, \quad \mu = 2.$$



Strong convexity and curvature

For a twice continuously differentiable function on $\mathbb{R}$,

$$f \text{ is } \mu\text{-strongly convex} \iff f''(x) \geq \mu \text{ for all } x.$$

Taylor's theorem turns the curvature bound into:

$$\begin{aligned} f(y) &= f(x) + f'(x)(y - x) + \frac{f''(\xi)}{2}(y - x)^2 \\ &\geq f(x) + f'(x)(y - x) + \frac{\mu}{2}(y - x)^2. \end{aligned}$$

Strict convexity and strong convexity

Strict convexity: for $x \neq y$ and $0 < t < 1$,

$$f((1-t)x + ty) < (1-t)f(x) + tf(y).$$

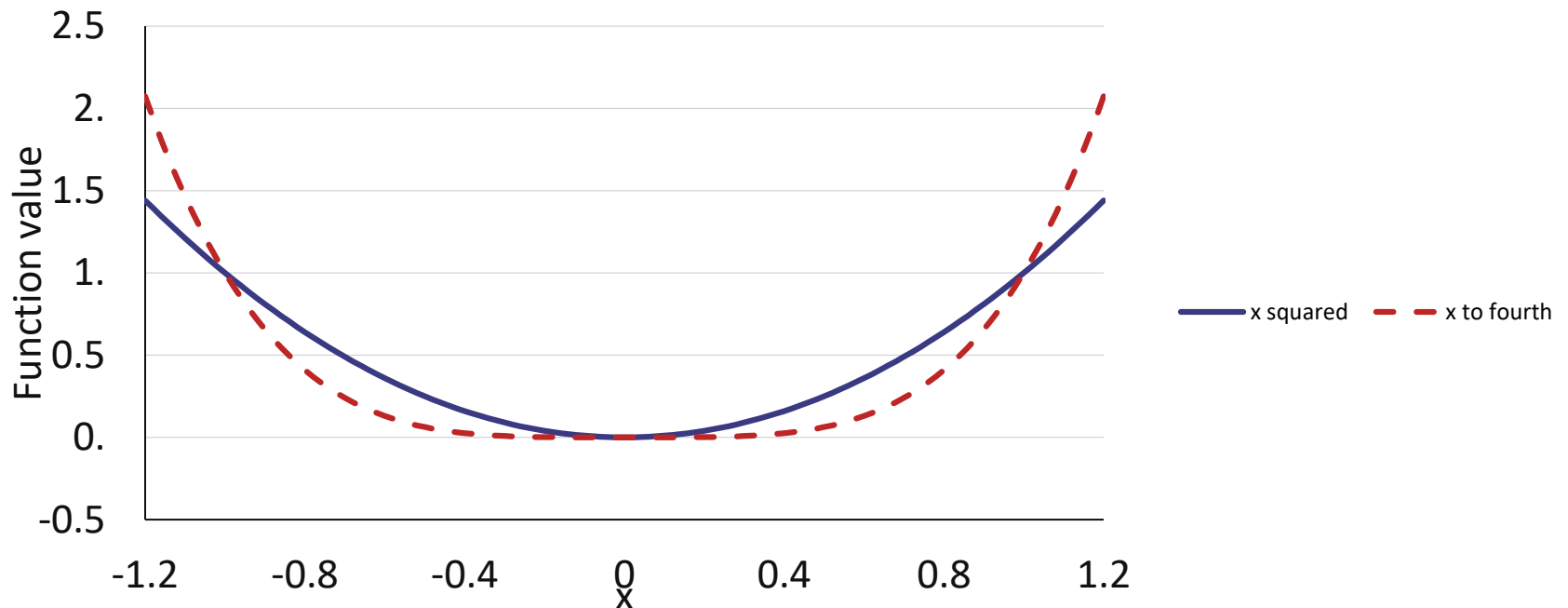
$$f(x) = x^2 : \quad f''(x) = 2 \quad (\mu = 2).$$

$$f(x) = x^4 : \quad f''(x) = 12x^2, \quad f''(0) = 0.$$

x^4 is strictly convex, but no positive global μ works.

Flatness near the minimum

$$f(x) = x^2, \quad g(x) = x^4.$$



Both have the unique minimizer 0. The curve x^4 becomes arbitrarily flat near its minimum.

Uniqueness of the minimizer

Let x^* be an unconstrained minimizer. Then $f'(x^*) = 0$.

$$f(x) \geq f(x^*) + \frac{\mu}{2}(x - x^*)^2.$$

If $x \neq x^*$, the added quadratic term is positive, so

$$f(x) > f(x^*).$$

The minimizer is **unique**.

Assumptions for the convergence proof

Assume f is μ -strongly convex and L -smooth on $\mathbb{R}$, with minimizer x^* .

$$0 < \mu \leq L, \quad x_{k+1} = x_k - \frac{1}{L} f'(x_k).$$

For a twice continuously differentiable function, these assumptions say:

$$\mu \leq f''(x) \leq L \quad \text{for every } x \in \mathbb{R}.$$

The derivative controls the value gap

Apply strong convexity at x , with $y = x^*$:

$$f(x^*) \geq f(x) + f'(x)(x^* - x) + \frac{\mu}{2}(x^* - x)^2.$$

We want to bound the value gap using only $f'(x)$.

The last two terms form a quadratic in $x^* - x$. **Complete the square.**

Completing the square

$$\begin{aligned} f(x^*) &\geq f(x) + f'(x)(x^* - x) + \frac{\mu}{2}(x^* - x)^2 \\ &= f(x) - \frac{[f'(x)]^2}{2\mu} + \frac{\mu}{2} \left(x^* - x + \frac{f'(x)}{\mu} \right)^2 \\ &\geq f(x) - \frac{[f'(x)]^2}{2\mu}. \end{aligned}$$

Rearranging gives:

$$\boxed{[f'(x)]^2 \geq 2\mu(f(x) - f(x^*))}.$$

One step of geometric decrease

Combine the descent inequality with the derivative bound:

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_k) - f(x^*) - \frac{[f'(x_k)]^2}{2L} \\ &\leq f(x_k) - f(x^*) - \frac{\mu}{L} (f(x_k) - f(x^*)) \\ &= \left(1 - \frac{\mu}{L}\right) (f(x_k) - f(x^*)). \end{aligned}$$

Each step removes at least a μ/L fraction of the current value gap.

The geometric convergence rate

Repeatedly applying the one-step bound gives, for $K \geq 1$,

$$f(x_K) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^K (f(x_0) - f(x^*)).$$

This is called **linear convergence**: the error decreases at a geometric rate.

If $\mu = L$, the bound gives the minimum after one step.

The condition number

$$\kappa = \frac{L}{\mu} \geq 1, \quad \text{contraction factor} = 1 - \frac{1}{\kappa}.$$

A large κ gives a small guaranteed reduction per step.

$$f(x_K) - f(x^*) \leq e^{-K/\kappa} (f(x_0) - f(x^*)).$$

To reduce the gap by a factor $\varepsilon \in (0, 1)$, it suffices to take $K \geq \kappa \log(1/\varepsilon)$.

Strong convexity in several variables

Replace products by inner products and squares by squared Euclidean norms:

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

For a twice continuously differentiable function, this means

$$v^T \nabla^2 f(x) v \geq \mu \|v\|^2 \quad \text{for all } x, v.$$

The same convergence proof

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k).$$

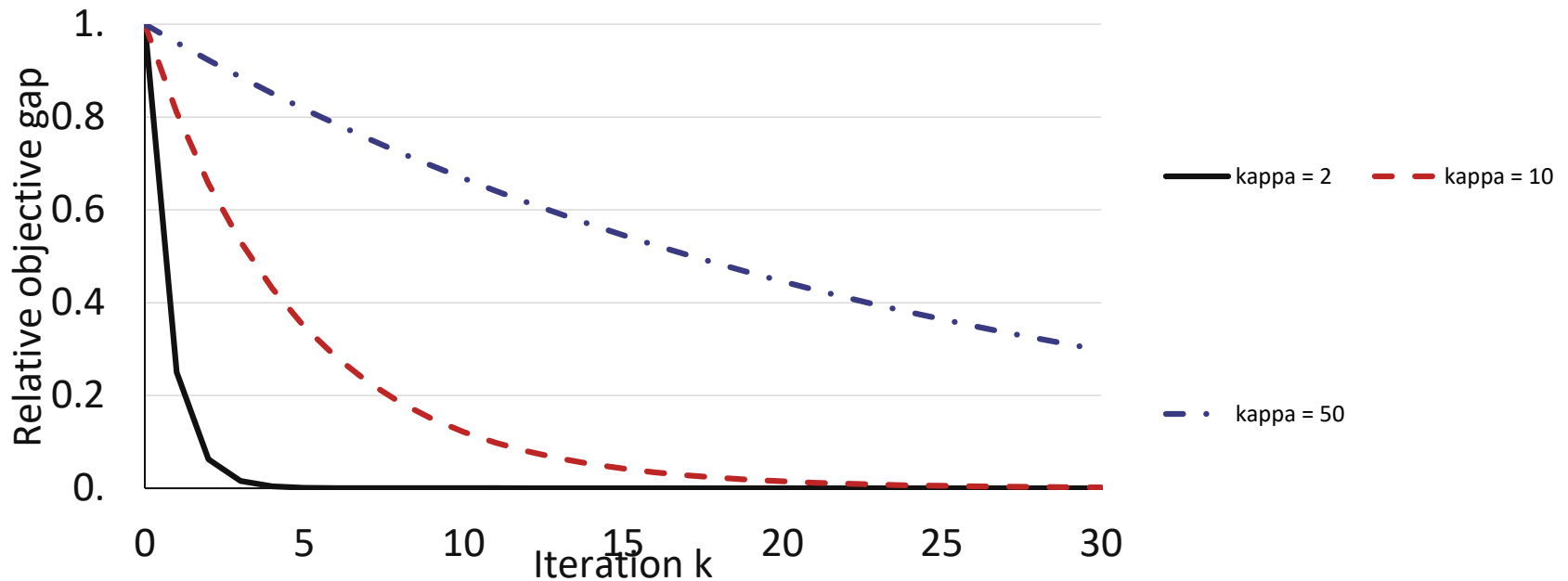
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x^*)).$$

$$f(x_K) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^K (f(x_0) - f(x^*)).$$

Assume global L -smoothness and μ -strong convexity. The same proof uses a completed squared norm.

Comparing different condition numbers

$$f(x_1, x_2) = \frac{1}{2}(x_1^2 + \kappa x_2^2), \quad x_0 = (1, 0), \quad \eta = 1/\kappa.$$



Actual relative gap: $\frac{f(x_k) - f(x^*)}{f(x_0) - f(x^*)} = \left(1 - \frac{1}{\kappa}\right)^{2k}.$

An objective with many terms

In data fitting, the objective is often an average of many losses:

$$f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x), \quad f'(x) = \frac{1}{N} \sum_{i=1}^N f'_i(x).$$

A full gradient step evaluates all N derivative terms.

Can we make a cheaper step using **one uniformly sampled term**?

Stochastic gradient descent

Draw a fresh index i_k uniformly from $\{1, \dots, N\}$, independently of earlier samples.

$$g_k = f'_{i_k}(x_k), \quad x_{k+1} = x_k - \eta_k g_k.$$

$$\mathbb{E}[g_k \mid x_k] = \frac{1}{N} \sum_{i=1}^N f'_i(x_k) = f'(x_k).$$

The estimate is **unbiased**, but an individual sample can give a poor direction.

Two losses with different minimizers

$$f_1(x) = \frac{1}{2}(x-1)^2, \quad f_2(x) = \frac{1}{2}(x+1)^2.$$

$$f(x) = \frac{f_1(x) + f_2(x)}{2} = \frac{x^2 + 1}{2}, \quad x^* = 0.$$

$$f'_1(0) = -1, \quad f'_2(0) = 1, \quad f'(0) = 0.$$

A sampled derivative can be nonzero at the average loss's minimum.

A stochastic step can increase the value

Start at $x_0 = 0$ and use $\eta = 1/2$. Each loss is equally likely.

$$x_1 = \begin{cases} \frac{1}{2}, & \text{if loss 1 is sampled,} \\ -\frac{1}{2}, & \text{if loss 2 is sampled.} \end{cases}$$

$$f(x_0) = \frac{1}{2}, \quad f(x_1) = \frac{5}{8} > \frac{1}{2}.$$

Unbiasedness does not imply descent at every step.

Reading

Chong, Lu, and Zak, *An Introduction to Optimization*, fifth edition.

Chapter 8, especially Sections 8.2–8.3: steepest descent and its analysis. Read alongside today's proof.

Review Section 22.2: convex functions. Use these slides for the strong-convexity argument.

Preview Section 27.1: the stochastic gradient descent algorithm.

Conceptual questions

1. Why can gradient descent still be slow for a strongly convex function?
2. Why can a stochastic gradient step increase the objective even when the gradient estimate is unbiased?