Convergence of gradient descent

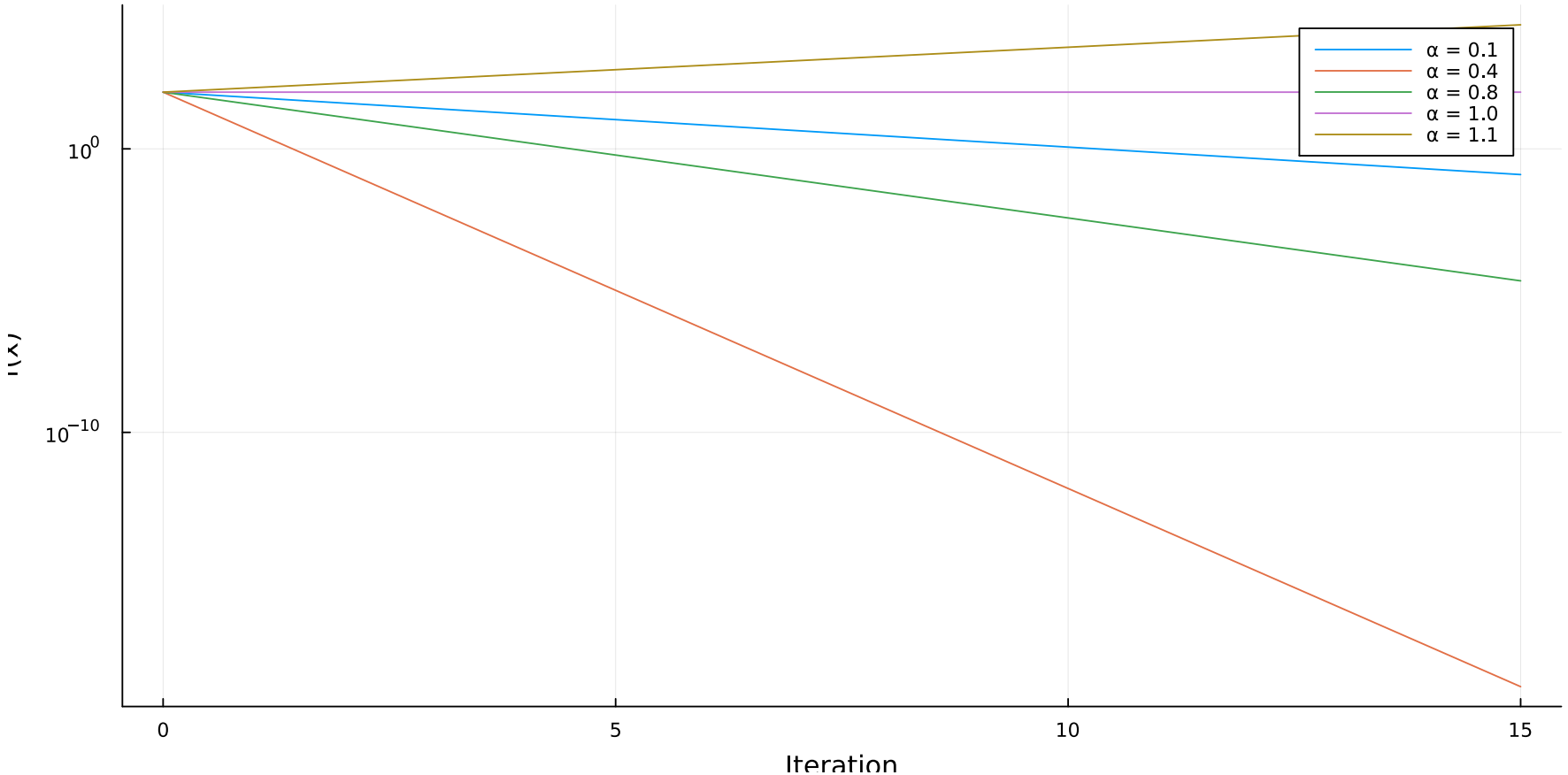
CPSC 406 – Computational Optimization

Convergence of gradient descent

- iteration complexity
- quadratic models
- descent lemma
- smoothness and strong convexity

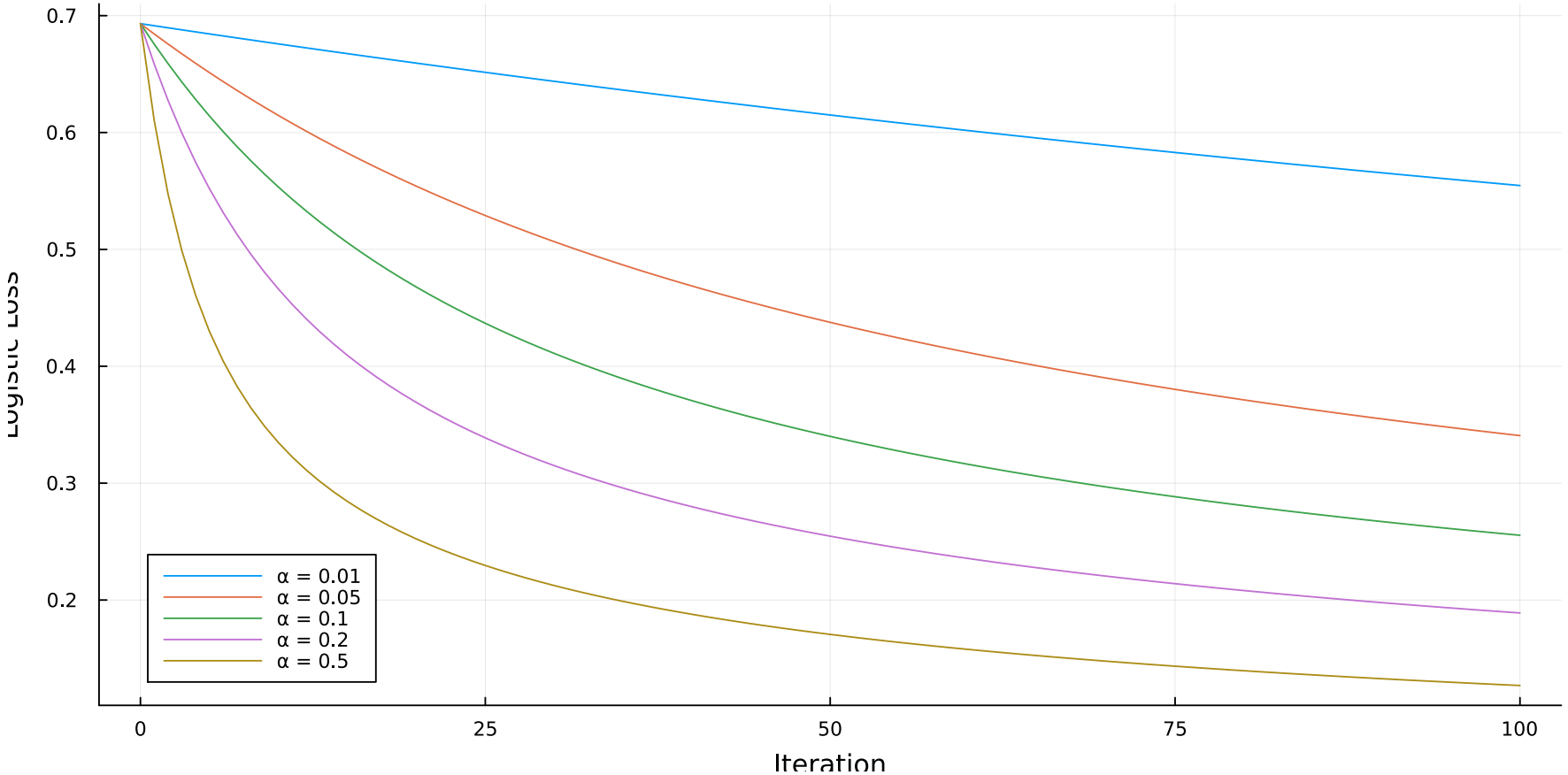
Example I: Step size selection

Gradient descent for $f(x) = x^2$ with different step sizes.



Example II: Logistic regression

Gradient descent for logistic regression $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i(x_i^T\theta)))$



Smooth functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth (ie, L-Lipschitz gradient)

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y$$

examples

- linear: $f(x) = a^T x$, with $a \in \mathbb{R}^n$, has $L = 0$
- quadratic: $f(x) = \frac{1}{2}x^T Ax + b^T x + \gamma$, with $A \succeq 0$, has $L = \|A\|_2 = \lambda_{\max}(A)$

Second-order characterization

If f is twice continuously differentiable, then f is L-smooth if and only if for all x

$$\nabla^2 f(x) \preceq LI \quad \text{ie,} \quad \|\nabla^2 f(x)\|_2 \leq L$$

Question

What is the Lipschitz constant L for the gradient of the function

$$f(x) = \frac{1}{2}\|cAx - b\|^2 \, ?$$

a. $L = c\,\|A\|^2$

b. $L = c^2\,\|A\|^2$

c. $L = \|A\|^2$

d. $L = \frac{\|A\|^2}{c}$

Descent lemma

If f is L-smooth, then for all x, z

$$f(z) \leq f(x) + \nabla f(x)^T (z - x) + \frac{L}{2} \|z - x\|^2$$

means that any L-smooth function is globally majorized by a quadratic approximation

Projected gradient descent

- projected gradient method for minimizing L-smooth f over a convex set C

$$x_{k+1} = \mathbf{proj}_C(x_k - \alpha \nabla f(x_k))$$

- by descent lemma, because $\alpha \leq 1/L$

$$f(z) \leq f(x) + \nabla f(x)^T (z - x) + \frac{L}{2}\|z - x\|^2 \leq f(x) + \nabla f(x)^T (z - x) + \frac{1}{2\alpha}\|z - x\|^2$$

- projected gradient descent (with step size α) step minimizes the **quadratic upper bound**:

$$\begin{aligned}
\mathbf{proj}_C(x - \alpha \nabla f(x)) &= \operatorname*{argmin}_{z \in C} \frac{1}{2\alpha}\|z - (x - \alpha \nabla f(x))\|^2 \\
&= \operatorname*{argmin}_{z \in C} \frac{\alpha}{2}\|\nabla f(x)\|^2 + \nabla f(x)^T (z - x) + \frac{1}{2\alpha}\|z - x\|^2 \\
&= \operatorname*{argmin}_{z \in C} f(x) + \nabla f(x)^T (z - x) + \frac{1}{2\alpha}\|z - x\|^2
\end{aligned}$$

Convergence

- Let $C = \mathbb{R}^n$ (unconstrained), $f_k := f(x_k), \nabla f_k := \nabla f(x_k)$. By descent lemma,

$$f_{k+1} \leq f_k + \nabla f_k^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

- take $x_{k+1} = x_k - \alpha \nabla f_k$, then $x_{k+1} - x_k = -\alpha \nabla f_k$ and

$$\begin{aligned} f_{k+1} &\leq f_k - \alpha \nabla f_k^T \nabla f_k + \frac{L}{2} \| - \alpha \nabla f_k\|^2 \\ &= f_k - \alpha \|\nabla f_k\|^2 + \frac{L\alpha^2}{2} \|\nabla f_k\|^2 \\ &= f_k - \alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla f_k\|^2 \end{aligned}$$

- decreasing objective values

$$f_{k+1} < f_k \quad \text{if} \quad \alpha \in (0, 2/L) \quad \text{and} \quad \nabla f_k \neq 0$$

Nonasymptotic rate

- if $\alpha \in (0, 2/L]$ then $f_{k+1} \leq f_k - \alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla f_k\|^2$
- minimize RHS over $\alpha \in (0, 2/L]$ gives $\alpha^* = 1/L$ and

$$f_{k+1} \leq f_k - \frac{1}{2L} \|\nabla f_k\|^2$$

- sum over $k = 0, 1, 2, \ldots, T$ gives

$$\frac{1}{2L} \sum_{k=0}^{T} \|\nabla f_k\|^2 \leq f(x_0) - f(x_T) \leq f(x_0) - f^*, \quad \text{where } f^* \text{ is min value}$$

- bounds min gradient value

$$\min_{k \in \{0,\ldots,T\}} \|\nabla f(x_k)\|^2 \leq \frac{1}{T} \sum_{k=0}^{T} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f^*)}{T} = O(1/T)$$

Question

Convergence rate of gradient descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex, L-smooth function. When applying gradient descent with a constant step size $\alpha = 1/L$, which of the following statements about the convergence is true?

a. The function values $f(x_k)$ decrease quadratically with the number of iterations k.

b. The gradient norms $\nabla f(x_k)$ converge to zero at a rate $O(1/k)$.

c. The method achieves a convergence rate of $O(e^{-k})$ for the function values.

d. The sequence $\{x_k\}$ generated converges to the minimizer in a finite number of steps.

Strong convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **μ-strongly convex** (with $\mu > 0$) if for all x, y

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

If f is **twice continuously differentiable**, then f is μ-strongly convex if and only if for all x

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2 \quad \forall d \in \mathbb{R}^n \qquad \iff \qquad \nabla^2 f(x) \succeq I\mu$$

Example (Quadratic functions) For a positive definite matrix A, the function

$$f(x) = \frac{1}{2} x^T A x + b^T x + \gamma$$

is μ-strongly convex with $\mu = \lambda_{\min}(A)$.

Alternative characterization

A function f is μ-strongly convex if and only if for all x

$$g(x) = f(x) - \frac{\mu}{2}\|x\|^2$$

is convex.

- Imples that **Tikhonov regularization** induces strong convexity

Distance to solution

Lemma 1 (Lipschitz smooth) If f is L-smooth, then for all x and all minimizers x^* with $f^* = f(x^*)$,

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f^* \leq \frac{L}{2}\|x - x^*\|^2$$

- gradient norm does not bound the distance to the solution

Lemma 2 (Strongly convex) If f is μ-strongly convex, then for all x and all minimizers x^* with $f^* = f(x^*)$,

$$\frac{\mu}{2}\|x - x^*\|^2 \leq f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2$$

Smoothness and strong convexity

- L smoothness imples

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2$$

- μ strong convexity implies

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

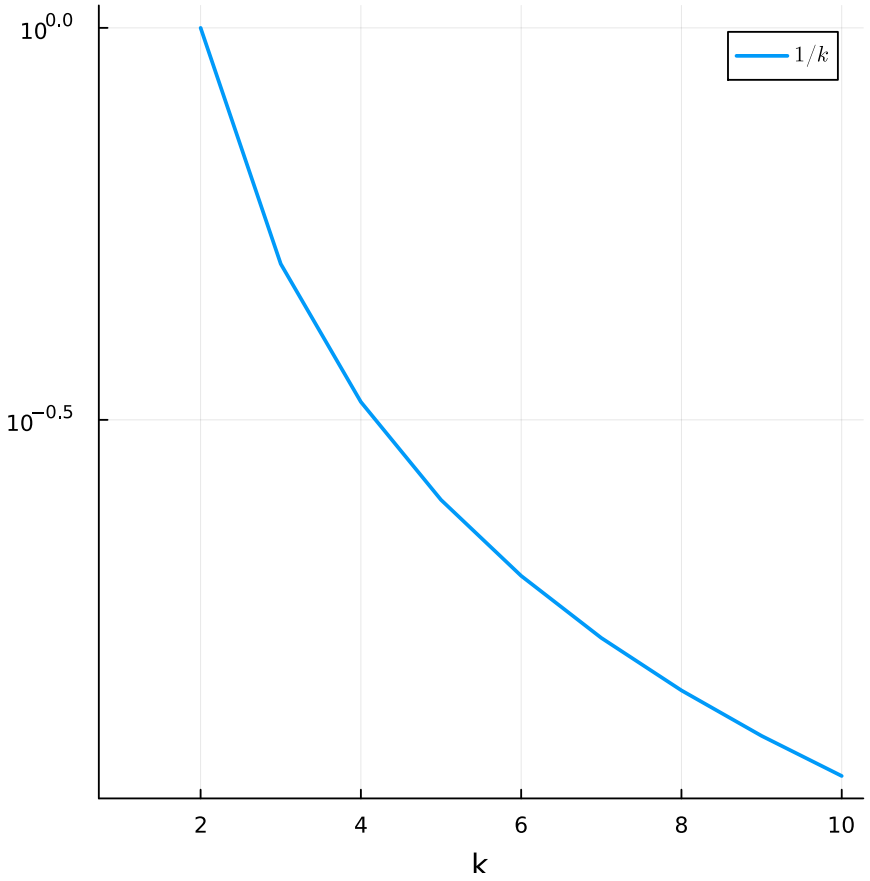
- together, for all x, y

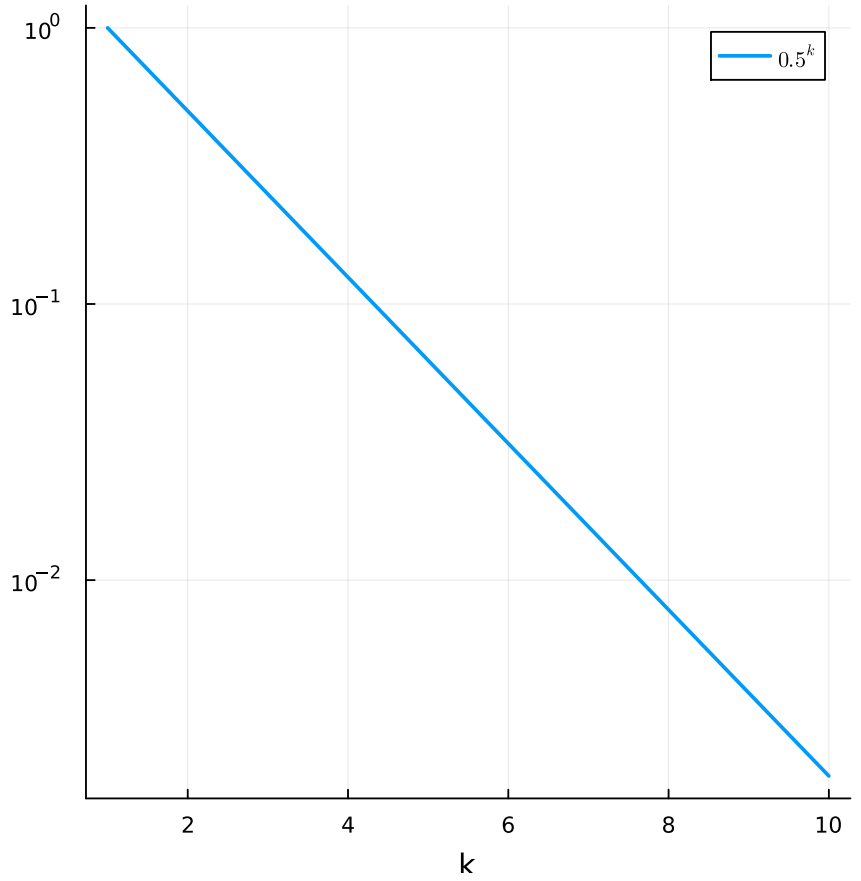
$$\frac{\mu}{2} \|y - x\|^2 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2} \|y - x\|^2$$

- implies Hessian eigenvalues bounded above and below:

$$\mu I \preceq \nabla^2 f(x) \preceq LI \quad \forall x$$

Linear convergence





Linear convergence with strong convexity

- under L-smoothness, we deduced the per-iteration decrease

$$f_{k+1} \leq f_k - \frac{1}{2L}\|\nabla f_k\|^2$$

- under μ-strong convexity, $\frac{1}{2\mu}\|\nabla f_k\|^2 \geq f_k - f^*$, hence

$$f_{k+1} \leq f_k - \frac{\mu}{L}(f_k - f^*) \quad \iff \quad f_{k+1} - f^* \leq (1 - \frac{\mu}{L})(f_k - f^*)$$

- recursing down from $k = T, T-1, \ldots, 0$ gives

$$f_T - f^* \leq (1 - \frac{\mu}{L})^T (f_0 - f^*) \leq \exp\left(-\frac{\mu}{L}T\right)(f_0 - f^*)$$

- if we require $f_T - f^* \leq \epsilon$, it's sufficient to run T iterations such that

$$T \geq \frac{L}{\mu}\log\left(\frac{f_0 - f^*}{\epsilon}\right) \quad \text{where } \frac{L}{\mu} \text{ is the condition number}$$