Steepest Descent $\begin{cases} x_{k+1} = x_k - \eta_k \nabla f(x_k) \\ \eta_k = \underset{\eta>0}{\operatorname{argmin}} f(x_k - \eta \nabla f(x_k)) \end{cases}$

Theorem 2.13. *For a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, assume $\mu I \le \nabla^2 f(x) \le LI$ where $L > \mu > 0$ are constants (eigenvalues of Hessian have uniform positive bounds), thus f is strongly convex has a unique minimizer $\mathbf{x}_*$. Then the steepest descent method (2.9) satisfies*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \le \left(1 - \frac{\mu}{L}\right)^k [f(\mathbf{x}_0) - f(\mathbf{x}_*)].$$

Remark: ① With strong convexity, and L-Cont. ∇f,

$$\|x_k - x^*\|^2 \le \left(1 - \frac{2\eta \mu L}{\mu + L}\right)^k \|x_0 - x^*\|^2$$

$$f(x_k) \le f(x_*) + \nabla f(x_*)^T (x_k - x_*) + \frac{L}{2}\|x_k - x_*\|^2$$

$$\Rightarrow \quad f(x_k) - f(x_*) \le \frac{L}{2}\|x_k - x_*\|^2$$

$$\eta = \frac{2}{L+\mu} \Rightarrow \left(\frac{L-\mu}{L+\mu}\right)^2 < 1 - \frac{\mu}{L}$$

$\Rightarrow$ provable rate of Steepest Descent is worse...

② For a quadratic cost function, the better rate $\left(\frac{L-\mu}{L+\mu}\right)^2$ can be proven for steepest descent.

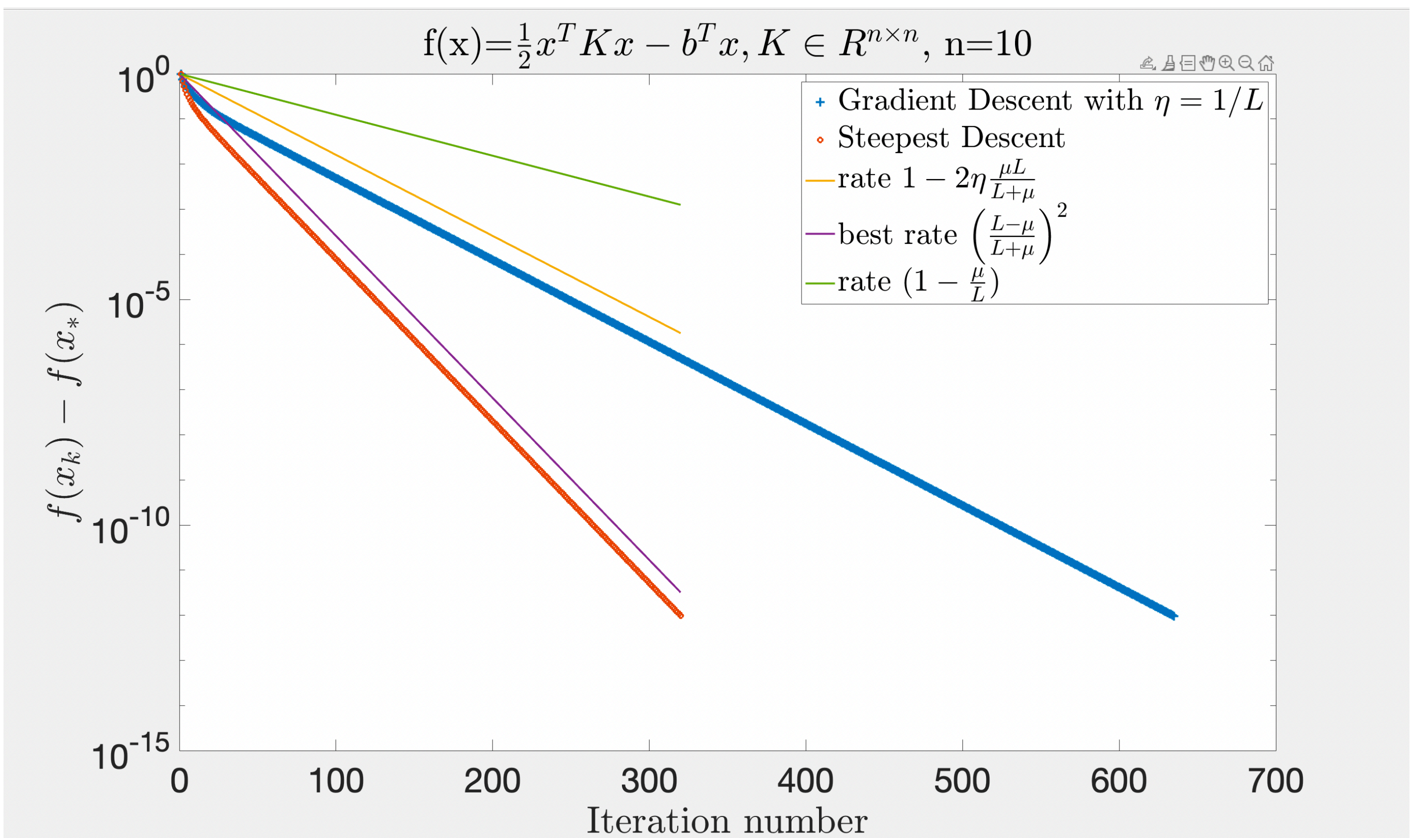
Numerical Example $\quad f(x) = \frac{1}{2} x^T K x - x^T b + C$

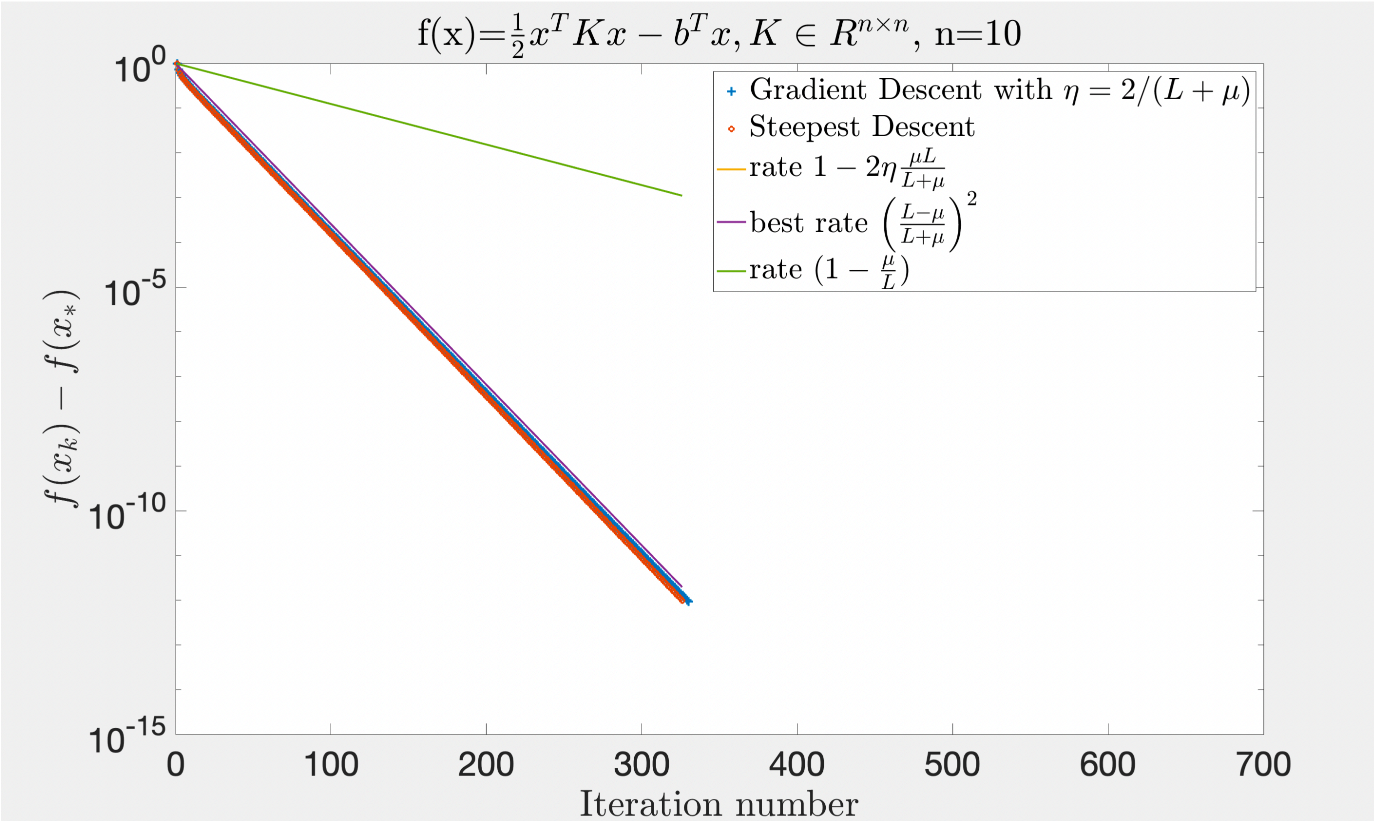
$$\nabla^2 f(x) = K \in \mathbb{R}^{n\times n} \qquad K = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \Delta x = \frac{1}{n+1}$$

$$\mu = \underset{\displaystyle \frac{4}{\Delta x^2}\sin^2\left(\frac{\pi}{2}\Delta x\right)}{\underset{\shortparallel}{\lambda_1(K)}} \le \cdots \le \underset{\displaystyle \frac{4}{\Delta x^2}\sin^2\left(\frac{\pi}{2}n\Delta x\right)}{\underset{\shortparallel}{\lambda_n(K)}} = L \Rightarrow \frac{L}{\mu} = O(n^2)$$

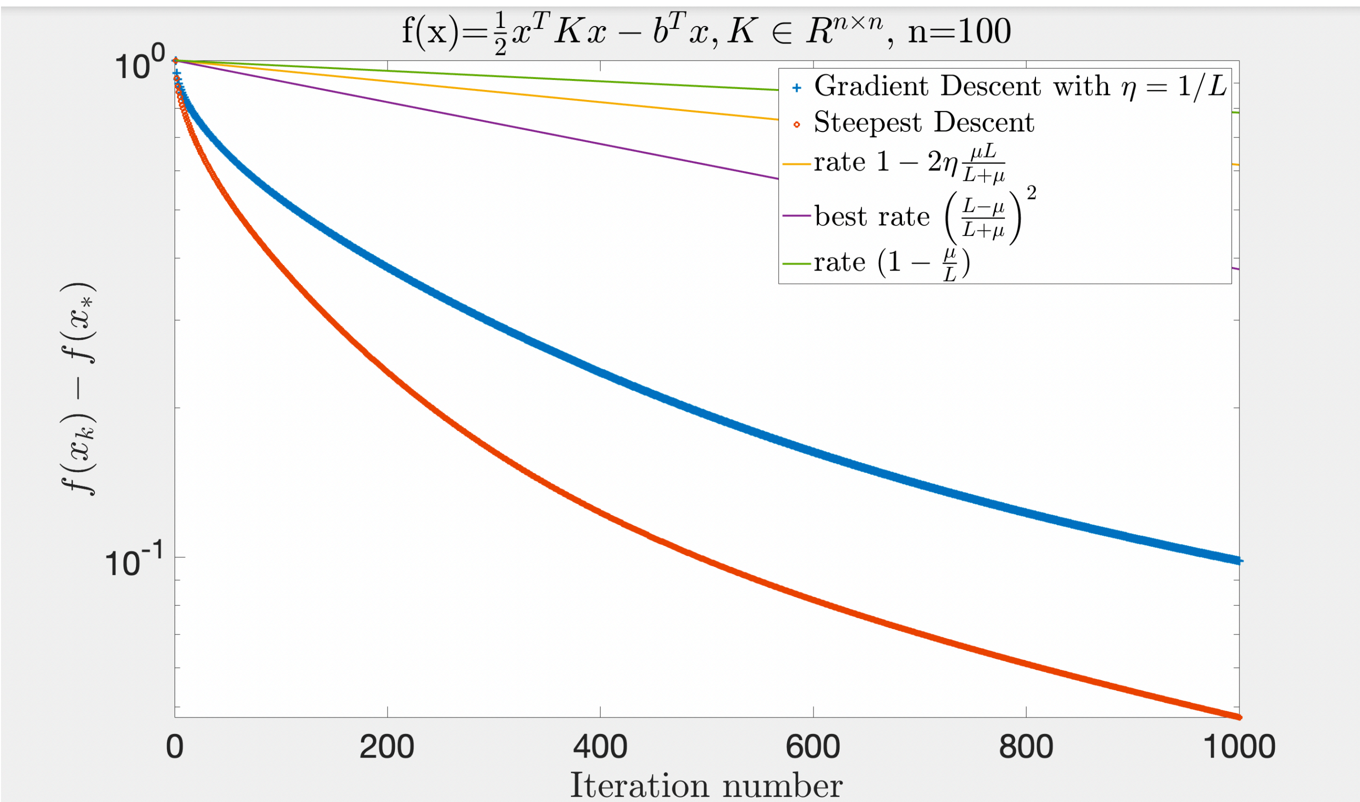
Provable Rates for $f(x_k) - f(x_*) = C^k$

GD with $\eta \le \frac{2}{L+\mu}$	$C = 1 - 2\eta\frac{\mu L}{L+\mu}$
$\eta = \frac{2}{L+\mu}$	$C = \left(\frac{L-\mu}{L+\mu}\right)^2$
Steepest Descent	$C = 1 - \frac{\mu}{L}$
Steepest Descent for quadratics	$C = \left(\frac{L-\mu}{L+\mu}\right)^2$

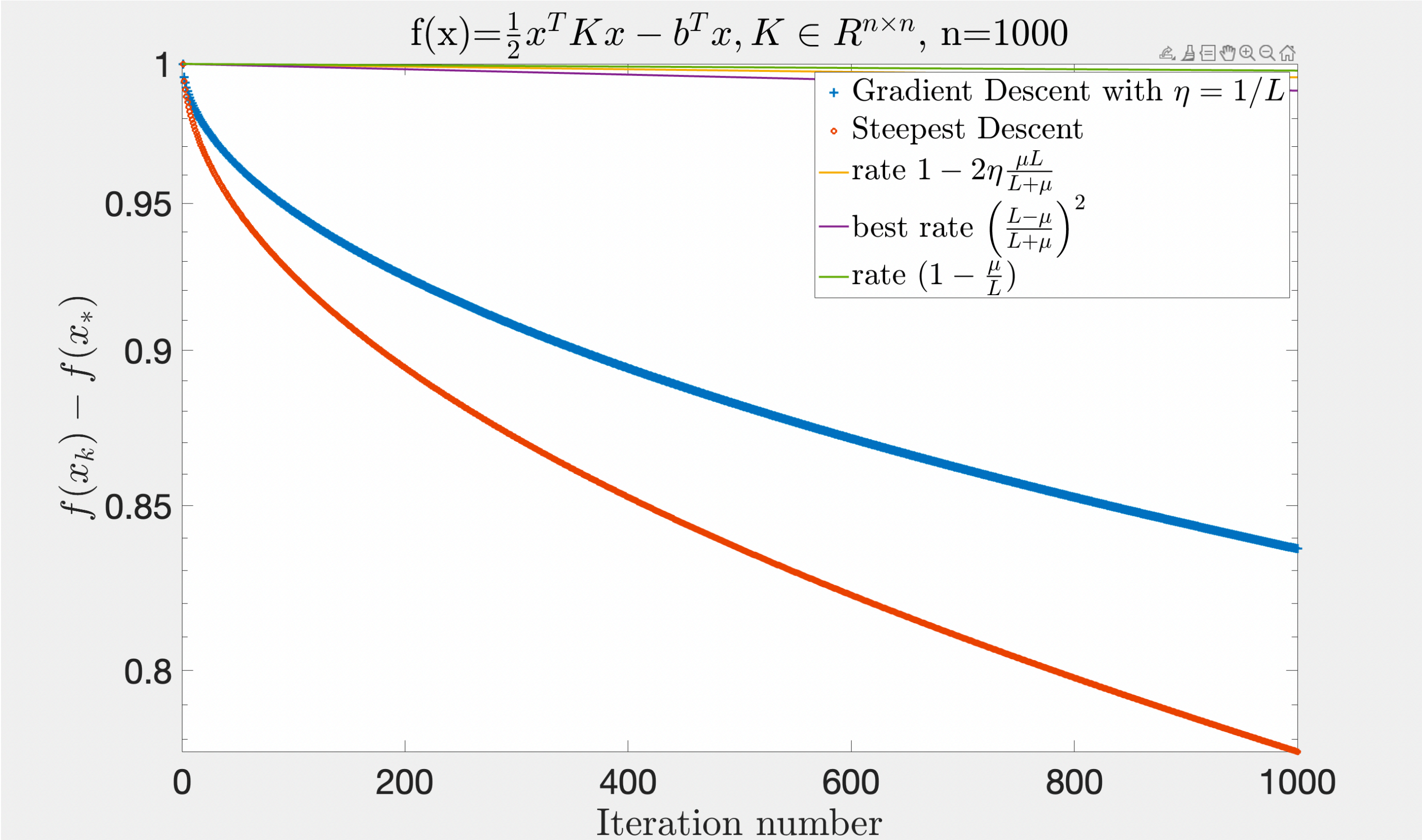




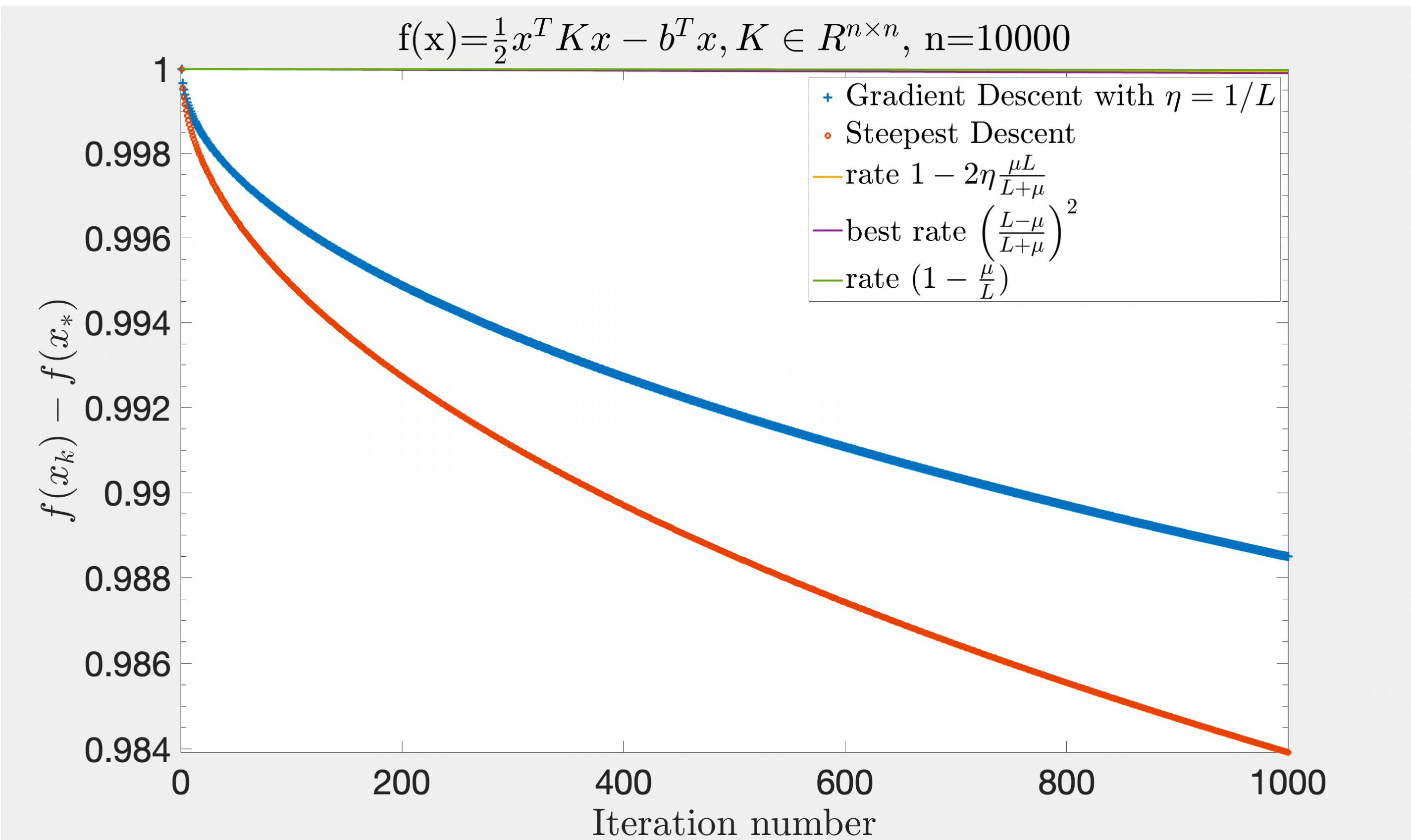
f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=10
10^0
10^{-5}
10^{-10}
10^{-15}
$f(x_k) - f(x_*)$
0
100
200
300
400
500
600
700
Iteration number
Gradient Descent with $\eta = 2/(L+\mu)$
Steepest Descent
rate $1 - 2\eta\frac{\mu L}{L+\mu}$
best rate $\left(\frac{L-\mu}{L+\mu}\right)^2$
rate $(1 - \frac{\mu}{L})$



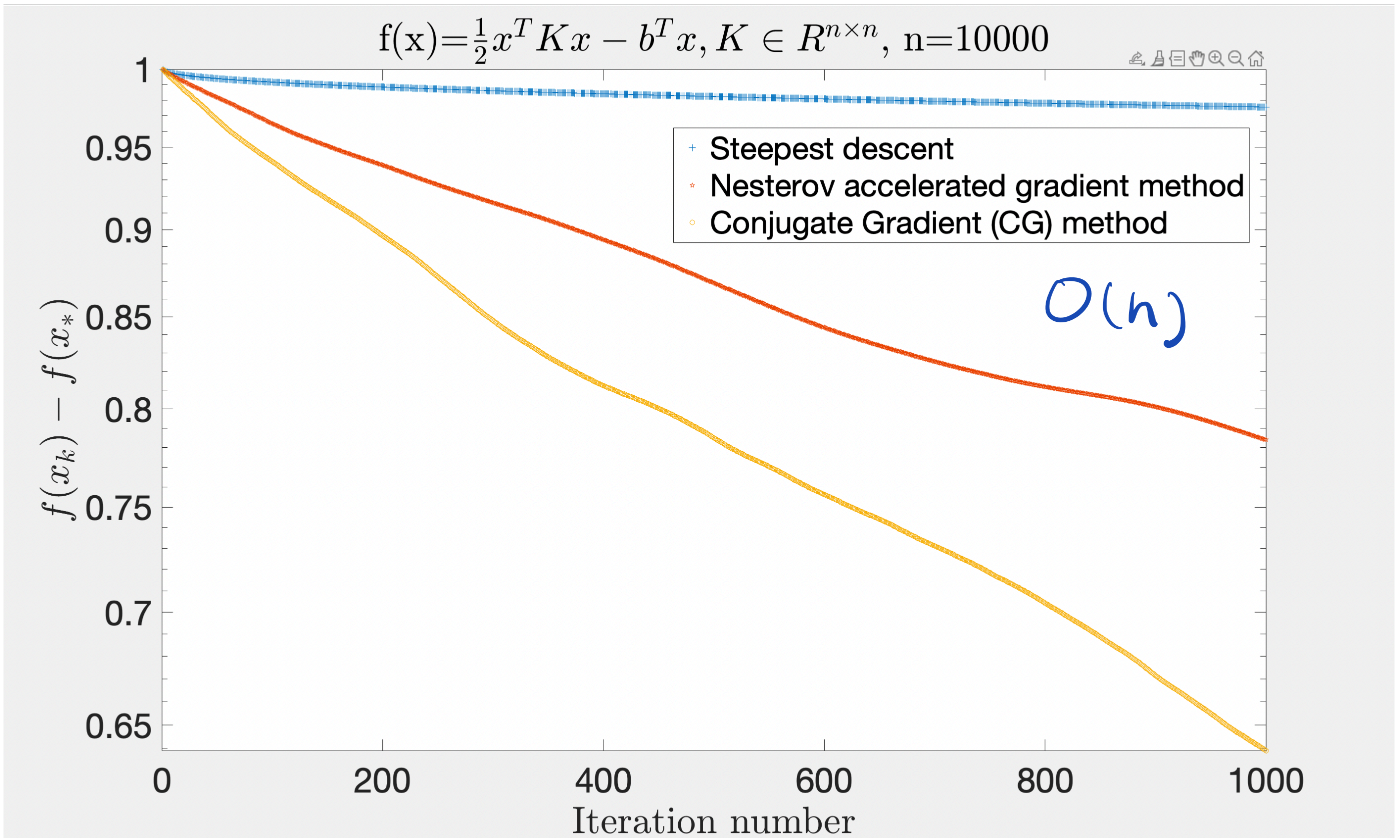
f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=100
10^0
10^{-1}
$f(x_k) - f(x_*)$
0
200
400
600
800
1000
Iteration number
Gradient Descent with $\eta = 1/L$
Steepest Descent
rate $1 - 2\eta\frac{\mu L}{L+\mu}$
best rate $\left(\frac{L-\mu}{L+\mu}\right)^2$
rate $(1 - \frac{\mu}{L})$



f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=1000
Gradient Descent with $\eta = 1/L$
Steepest Descent
rate $1 - 2\eta\frac{\mu L}{L+\mu}$
best rate $\left(\frac{L-\mu}{L+\mu}\right)^2$
rate $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
1
0.95
0.9
0.85
0.8
0
200
400
600
800
1000
Iteration number



f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=10000
Gradient Descent with $\eta = 1/L$
Steepest Descent
rate $1 - 2\eta\frac{\mu L}{L+\mu}$
best rate $\left(\frac{L-\mu}{L+\mu}\right)^2$
rate $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
1
0.998
0.996
0.994
0.992
0.99
0.988
0.986
0.984
0
200
400
600
800
1000
Iteration number



Nesterov accelerated gradient method

$$\boxed{\begin{cases} \mathbf{x}_{k+1} &= \mathbf{y}_k - \eta_k \nabla f(\mathbf{y}_k) \\ t_{k+1} &= \frac{1}{2}\left(1 + \sqrt{4t_k^2 + 1}\right) \\ \mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) \end{cases} \qquad \mathbf{x}_0 = \mathbf{y}_0, t_0 = 1.}$$

2.3 Line search method

Now we consider a more general method for minimizing $f(\mathbf{x})$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{p}_k,$$

where $\eta_k > 0$ is a step size and $\mathbf{p}_k \in \mathbb{R}^n$ is a search direction. Examples of the search direction include:

1. *Gradient method* $\quad \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. *Newton's method* $\quad \mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1}\nabla f(\mathbf{x}_k)$. L-BFGS
3. *Quasi Newton's method* $\quad \mathbf{p}_k = -B_k \nabla f(\mathbf{x}_k)$, where $B_k \approx [\nabla^2 f(\mathbf{x}_k)]^{-1}$.
4. *Conjugate Gradient Method* $\quad \mathbf{p}_k = -(\mathbf{x}_k - \mathbf{x}_{k-1} + \beta_k \nabla f(\mathbf{x}_k))$, where β_k is designed such that $\mathbf{p}_k$ and $\mathbf{x}_k - \mathbf{x}_{k-1}$ are conjugate (orthogonal in some sense).

The search direction $\mathbf{p}_k$ is a descent direction if $\langle \mathbf{p}_k, -\nabla f(\mathbf{x}_k)\rangle > 0$, i.e., $\mathbf{p}_k$ pointing to the negative gradient direction.

2.3.1 The step size

To find a proper step size η_k, it is natural to ask for a sufficient decrease in the cost function:

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle, \quad c_1 \in (0, 1). \tag{2.12a}$$

The constant c_1 is usually taken as a small number such as 10^{-4}, and (2.12a) is called *Amijo condition.* To avoid unacceptably small step sizes, the *curvature condition* requires

$$\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k \rangle \geq c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle, \quad c_2 \in (c_1, 1). \tag{2.12b}$$

Define $\phi(\eta) = f(\mathbf{x}_k + \eta \mathbf{p}_k)$, then $\phi'(\eta) = \langle \nabla f(\mathbf{x}_k + \eta \mathbf{p}_k), \mathbf{p}_k \rangle$, thus (2.12b) simply requires $\phi'(\eta_k) \geq c_2 \phi'(0)$, where $\phi'(0) = \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle < 0$ for a descent direction $\mathbf{p}_k$. Usually, c_2 is taken as 0.9 for Newton and quasi-Newton methods, and 0.1 in conjugate gradient methods.

The two conditions in (2.12) with $0 < c_1 < c_2 < 1$ are called the *Wolfe conditions.*

The following are called the *strong Wolfe conditions.*

$$f(\mathbf{x}_k + \eta \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \eta \langle \nabla f(\mathbf{x}_k), p_k \rangle, \quad c_1 \in (0, 1). \tag{2.13a}$$

$$|\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k \rangle| \leq c_2 |\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle|, \quad c_2 \in (c_1, 1). \tag{2.13b}$$

Lemma 2.4. *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is continuously differentiable and has a lower bound, and $\mathbf{p}_k$ is a descent direction. Then for any $0 < c_1 < c_2 < 1$, there are intervals of η satisfying the Wolfe conditions* (2.12) *and the strong Wolfe conditions* (2.13).

$\langle p, -\nabla f \rangle > 0$

2.3.2 The convergence

We consider the angle θ_k between the negative gradient and the search direction:

$$\cos \theta_k = \frac{\langle -\nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle}{\|\nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\|}.$$

Theorem 2.16 (Zoutendijk's Theorem). *Assume $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is continuously differentiable with Lipschitz continuous gradient $\nabla f(\mathbf{x})$, and $f(\mathbf{x})$ is bounded from below. Consider a line search method $\mathbf{x}_{k+1} = \mathbf{x}_k + \eta_k \mathbf{p}_k$, where $\mathbf{p}_k$ is a descent direction and η_k satisfies the Wolfe conditions (2.12). Then*

$$\sum_{k=1}^{\infty} \cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\|^2 < +\infty.$$

Proof. By (2.12b), we have

$$\langle \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle \geq (c_2 - 1)\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle.$$

The Lipschitz continuity and Cauchy Schwartz inequality give

$$\langle \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle \leq \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\| \leq L \|\eta_k \mathbf{p}_k\| \|\mathbf{p}_k\|.$$

Combining the two inequalities, we get

$$\frac{\langle \nabla f, p \rangle}{\|p\|^2} = \underbrace{\frac{\langle \nabla f, p \rangle}{\|\nabla f\| \cdot \|p\|}}_{\cos\theta} \cdot \frac{\|\nabla f\|}{\|p\|}$$

$$\eta_k \geq \frac{c_2 - 1}{L} \frac{\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle}{\|\mathbf{p}_k\|^2}.$$

Plugging it into (2.12a), we get

$f(x_k + \eta p_k) \leq f(x_k) + c_1 \eta \langle \nabla f(x_k), p_k \rangle$

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \leq f(\mathbf{x}_k) - c_1 \frac{1 - c_2}{L} \frac{|\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle|^2}{\|\mathbf{p}_k\|^2},$$

which can be written as

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \omega \cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\|^2, \quad \omega = c_1 \frac{1 - c_2}{L}.$$

Summing it up, since $f(\mathbf{x}) \geq C$, we get

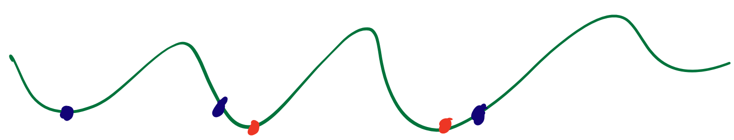
$$\sum_{k=0}^{N} \cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{1}{\omega}[f(\mathbf{x}_0) - f(\mathbf{x}_{N+1})] \leq \frac{1}{\omega}[f(\mathbf{x}_0) - C].$$

So $a_N = \sum_{k=0}^{N} \cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\|^2$ is a bounded and increasing sequence, thus the infinite series converges. ☐

The convergence of the series in Zoutendijk's Theorem gives $\cos^2 \theta_k \|\nabla f(\mathbf{x}_k)\| \to 0$. Thus if $\cos^2 \theta_k \geq \delta > 0, \forall k$, then $\|\nabla f(\mathbf{x}_k)\| \to 0$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$\sqrt{2}$



$f(x)$

Example 2.8. *Consider Newton's method with* $\mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1}\nabla f(\mathbf{x}_k)$*. Assume the Hessian has some uniform positive bounds for eigenvalues (i.e., the Hessian is* **positive definite** *with a uniformly bounded condition number:):*

$$\mu I \leq \nabla^2 f(\mathbf{x}) \leq LI, \quad L \geq \mu > 0, \forall \mathbf{x},$$

then we have (eigenvalues of A are reciprocals of eigenvalues of A^{-1})

$$\frac{1}{L} I \leq [\nabla^2 f(\mathbf{x})]^{-1} \leq \frac{1}{\mu} I, \quad L \geq \mu > 0, \forall \mathbf{x}.$$

For convenience, let $B_k = [\nabla^2 f(\mathbf{x})]^{-1}$ and $\mathbf{h}_k = \nabla f(\mathbf{x}_k)$. Since B_k is positive definite, its eigenvalues are also singular values. By the definition of spectral norm, we get $\|Ax\| \leq \|A\| \cdot \|x\|$ $\quad \|A\| = \max_x \frac{\|Ax\|}{\|x\|} = \max_i \sigma_i(A)$

$$\|\mathbf{p}_k\| = \|B_k \nabla f(\mathbf{x}_k)\| \leq \|B_k\| \|\nabla f(\mathbf{x}_k)\| \leq \frac{1}{\mu}\|\nabla f(\mathbf{x}_k)\| = \frac{1}{\mu}\|\mathbf{h}_k\|.$$

By the Courant-Fischer-Weyl min-max principle (Appendix A.1), we have

$$\cos\theta_k = \frac{\langle -\nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle}{\|\nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\|} = \frac{\mathbf{h}_k^T B_k \mathbf{h}_k}{\|\mathbf{h}_k\| \|\mathbf{p}_k\|} \geq \mu \frac{\mathbf{h}_k^T B_k \mathbf{h}_k}{\|\mathbf{h}_k\| \|\mathbf{h}_k\|} \geq \frac{\mu}{L} = \frac{1}{L/\mu},$$

where $L/\mu = \|B_k\| \|B_k^{-1}\|$ is the condition number of the Hessian. With Theorem 2.16, we get $\|\nabla f(\mathbf{x}_k)\| \to 0$. Recall that a strongly convex function has a unique critical point which is the global minimizer. So the Newton's method with a step size satisfying the Wolfe conditions (2.12) converges to the unique minimizer $\mathbf{x}_*$ for a strongly convex function $f(\mathbf{x})$ if $\|\nabla^2 f(\mathbf{x})\|$ has a uniform upper bound, see the problem below.

Problem 2.1. *Recall that $\|\nabla f(\mathbf{x}_k)\| \to 0$ may not even imply $\mathbf{x}_k$ converges to a critical point, see Example 2.2. Prove that $\|\nabla f(\mathbf{x}_k)\| \to 0$ implies $\mathbf{x}_k$ converges to the global minimizer under the assumption*

$$\mu I \leq \nabla^2 f(\mathbf{x}) \leq LI, \quad L \geq \mu > 0, \forall \mathbf{x}.$$