Convergence Rates of GD

Always assume $\nabla f(x)$ is Lip-cont. with L

$\omega = \eta(\frac{2}{L} - \eta)$

① No other assumptions, $\eta \in (0, \frac{2}{L})$

$$\min_{0 \le k \le N} \|\nabla f(x_k)\| \le \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}[f(x_0) - f(x_*)]} = O(\frac{1}{\sqrt{k}})$$

② $f(x)$ is convex, $\eta \in (0, \frac{2}{L})$

$$f(x_k) - f(x_*) < \frac{\|x_0 - x_*\|^2}{\omega} \frac{1}{k} = O(\frac{1}{k})$$

③ $f(x)$ is strongly convex with $\mu > 0$, $\eta \in (0, \frac{2}{L+\mu}]$

$$\|x_k - x^*\| \le c^k \|x_0 - x^*\| = O(c^k)$$

$$f(x_k) - f(x_*) \le c^{2k} \|x_0 - x^*\|^2 \quad c = \sqrt{1 - \frac{2\eta\mu L}{\mu + L}}$$

④ Polyak-Lojasiewicz (PL) inequality

$\eta = \frac{1}{L}$

(PL) $$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \ge \mu(f(\mathbf{x}) - f(\mathbf{x}_*)), \quad \mu > 0$$

$$f(x_k) - f(x_k) = \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f(x_*)]$$

1) Strong Convexity $\Rightarrow$ PL inequality

Strong Convexity $\Leftrightarrow$

1. $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x},$
2. $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y}.$

$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2$

$\Rightarrow \min_y f(y) \geq \min_y \left[f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \right]$

$0 = \partial_y [\ldots] = \nabla f(x) + \mu(y - x)$

$\Rightarrow$ minimizer y_* satisfies

$\mu(y_* - x) = -\nabla f(x)$

$\Rightarrow f(x_*) \geq f(x) + \langle \nabla f(x), \frac{\nabla f(x)}{-\mu} \rangle + \frac{\mu}{2}\left\|\frac{\nabla f(x)}{-\mu}\right\|^2$

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

2) $f(x) = \frac{1}{2}\|Ax - b\|^2$ may not be strongly convex but it satisfies (PL)

$\nabla^2 f = A^T A = [\square] \geq 0$

Example 2.22. *The logistic regression cost function*

$$f(\mathbf{x}) = \sum_{i=1}^{n} \log[1 + \exp(\mathbf{b}_i \mathbf{a}_i^T \mathbf{x})]$$

can be written as $f(\mathbf{x}) = g(A\mathbf{x})$, *with* $g(\mathbf{y})$ *being only strictly convex but not strongly convex. For example* $g(x) = \log(1 + \exp x)$ *has* $g''(x) = \frac{e^x}{(e^x+1)^2} \to 0$ *as* $x \to \infty$. *So Theorem 2.23 below does not apply but* $f(\mathbf{x})$ *satisfies the Polyak-Lojasiewicz inequality on a compact set.*

Theorem 2.23. *$f(\mathbf{x}) = g(A\mathbf{x})$ with $A \in \mathbb{R}^{m\times n}$ satisfies the Polyak-Lojasiewicz inequality if $g(\mathbf{x})$ is strongly convex.*

The Polyak-Lojasiewicz (PL) inequality or condition is given as

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f(\mathbf{x}_*)), \quad \mu > 0, \tag{2.9}$$

Theorem 2.24. *Let $f(\mathbf{x})$ satisfy the Polyak-Lojasiewicz inequality (2.9) with $\mu > 0$ and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L. The gradient method with a step-size of $1/L$, $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ has a global linear convergence rate*

$$f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_0) - f(\mathbf{x}_*)).$$

Remark 2.25. *Once again, $\mathbf{x}_k \to \mathbf{x}_*$ cannot be true in general since $\mathbf{x}_*$ may not be unique.*

Proof. The Descent Lemma (Lemma 2.1) gives

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

which becomes the following after using $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ and (2.9) ,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - \frac{\mu}{L}(f(\mathbf{x}_k) - f(\mathbf{x}_*)),$$

$$\Rightarrow f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq (1 - \frac{\mu}{L})(f(\mathbf{x}_k) - f(\mathbf{x}_*)).$$

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

Theorem 2.26. *For a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, assume $\mu I \le \nabla^2 f(\mathbf{x}) \le LI$ where $L > \mu > 0$ are constants (eigenvalues of Hessian have uniform positive bounds), thus $f(\mathbf{x})$ is strongly convex and has a unique minimizer $\mathbf{x}_*$. Then the steepest descent method (2.10) satisfies*

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

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 $f(x) = \frac{1}{2}x^T K x - x^T b + c$

$\nabla^2 f(x) = K \in \mathbb{R}^{n \times n}$

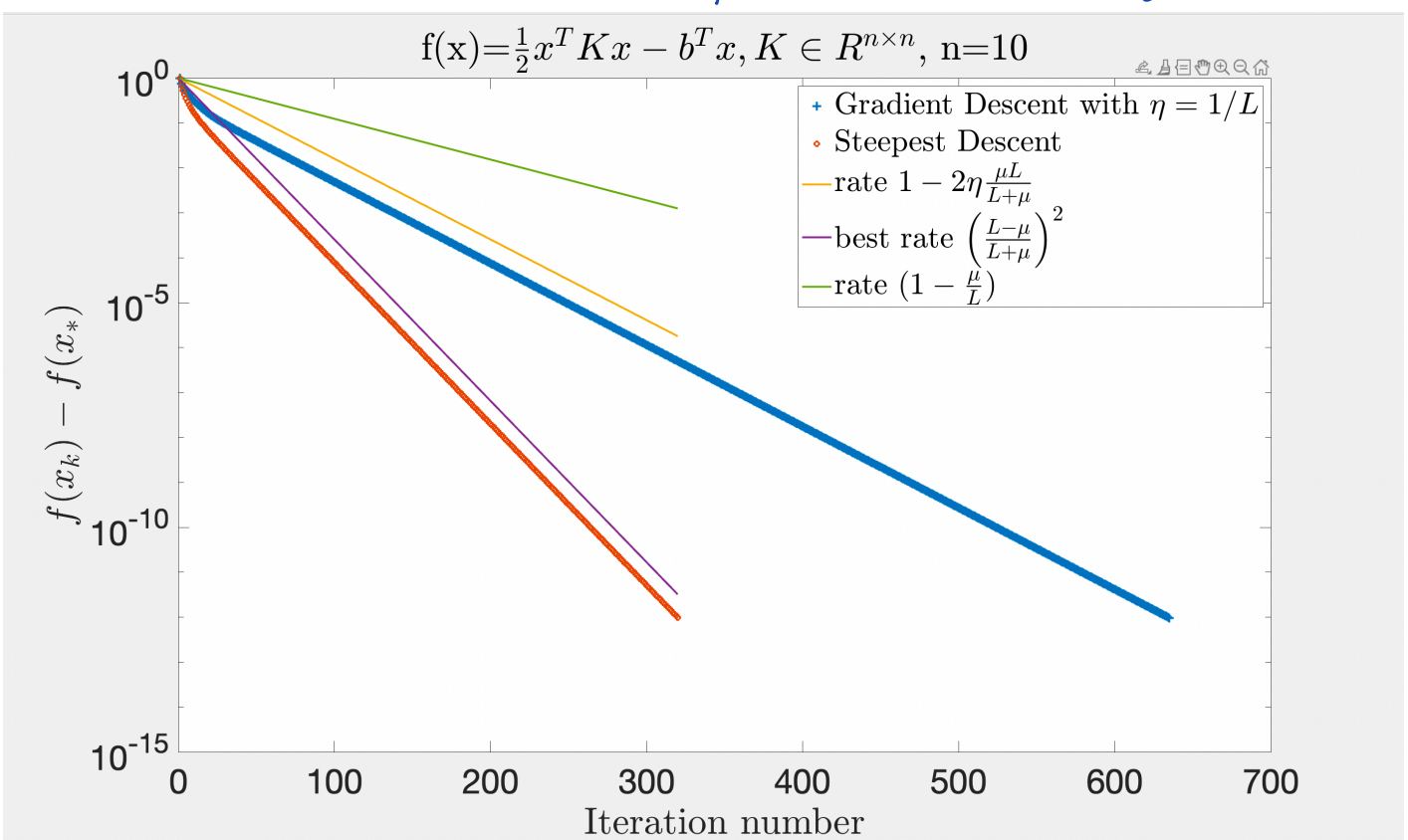
$K = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$, $\Delta x = \frac{1}{n+1}$

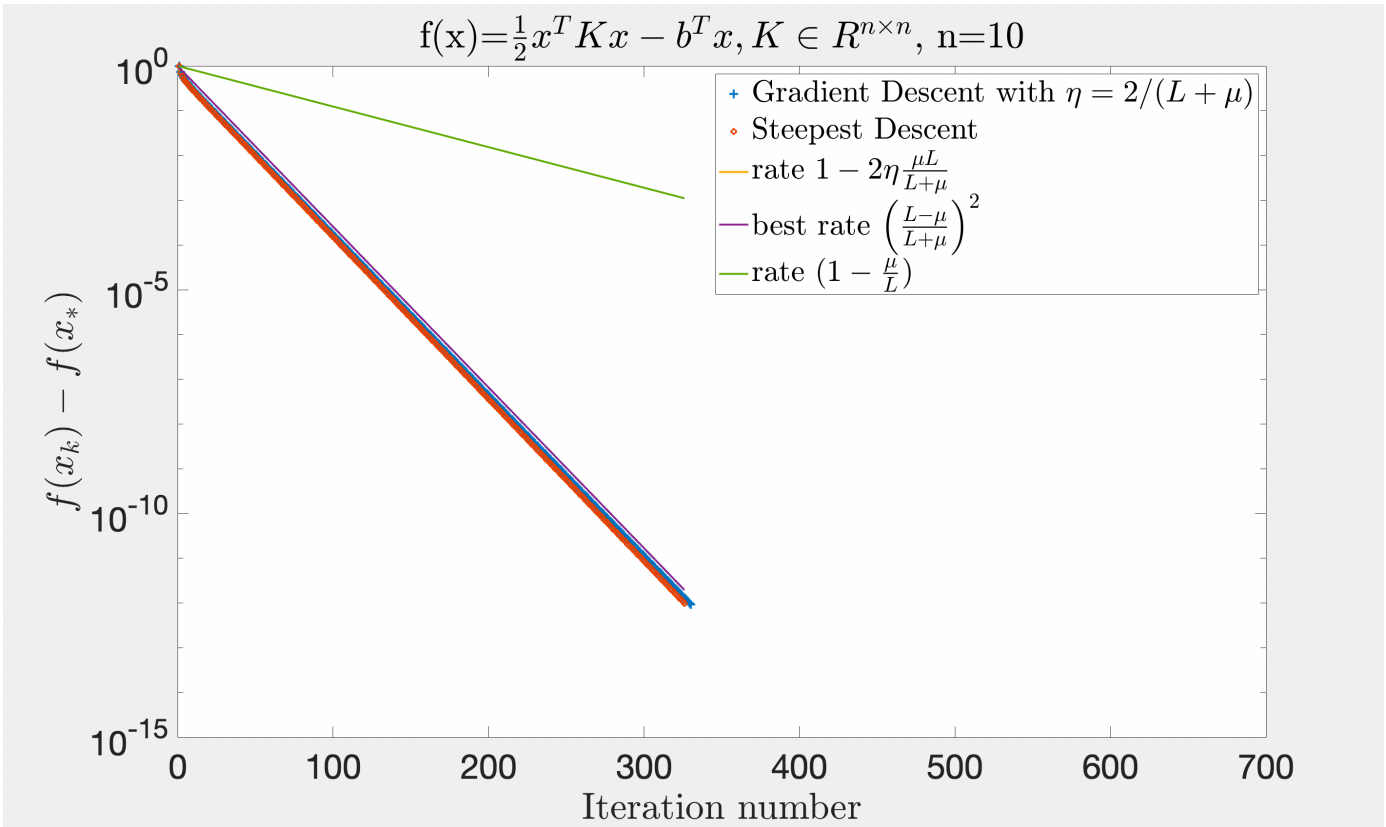
$\mu = \lambda_1(K) \leq \cdots \leq \lambda_n(K) = L$ (Appendix B)

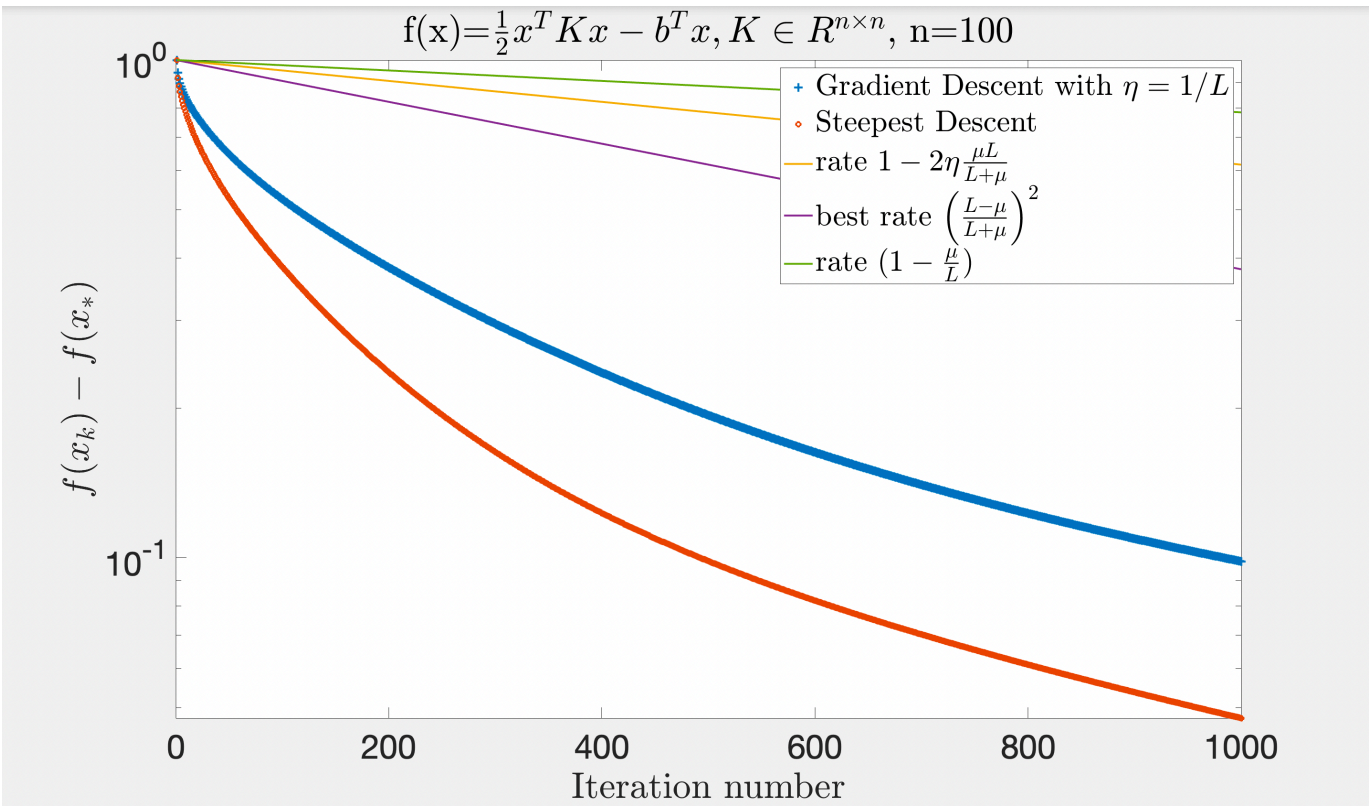
$\lambda_1 = \frac{4}{\Delta x^2}\sin^2\left(\frac{\pi}{2}\Delta x\right)$ $\quad$ $\lambda_n = \frac{4}{\Delta x^2}\sin^2\left(\frac{\pi}{2} n \Delta x\right) \Rightarrow \frac{L}{\mu} = O(n^2)$

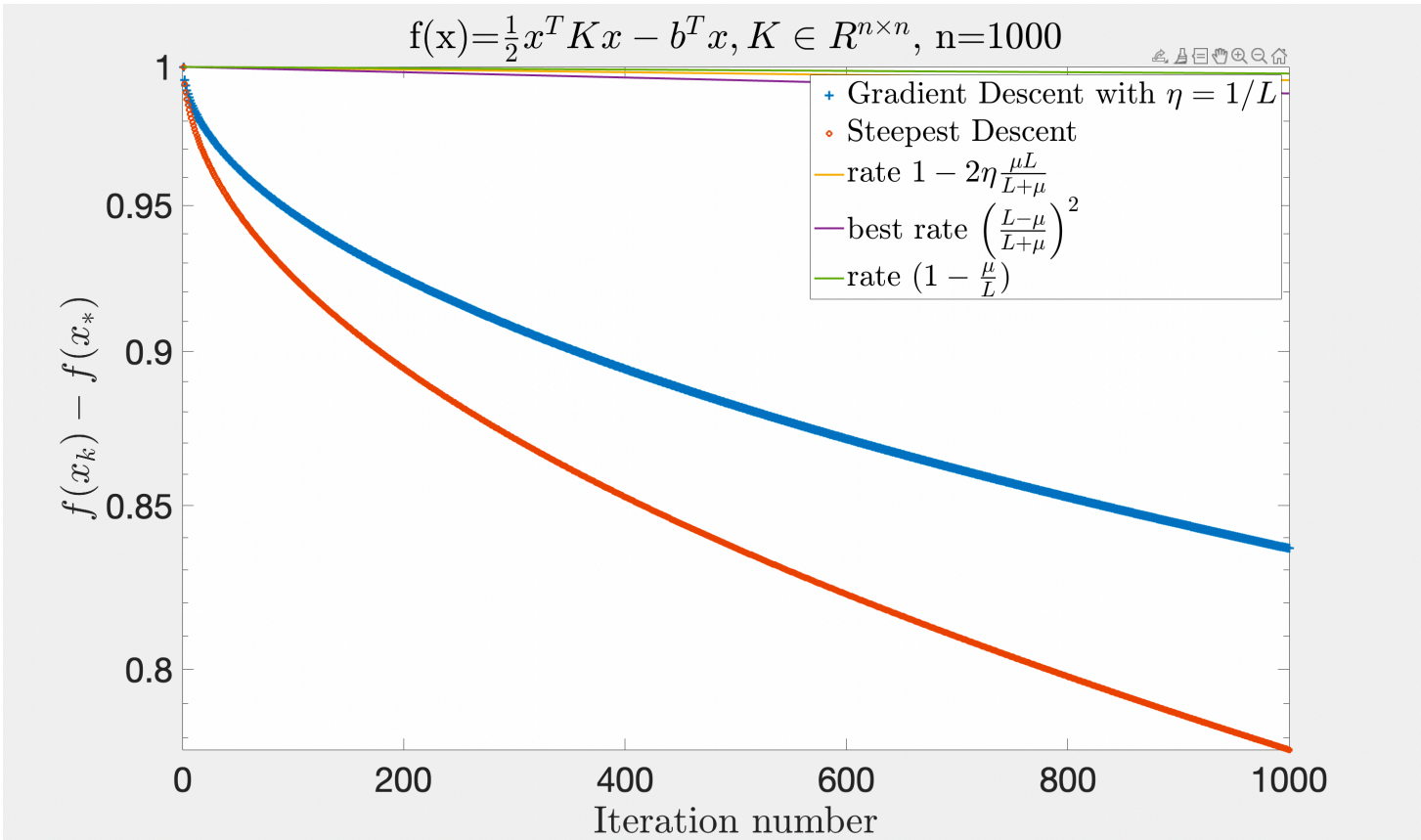
Provable Rates for $f(x_k) - f(x_*) \leq C^k$

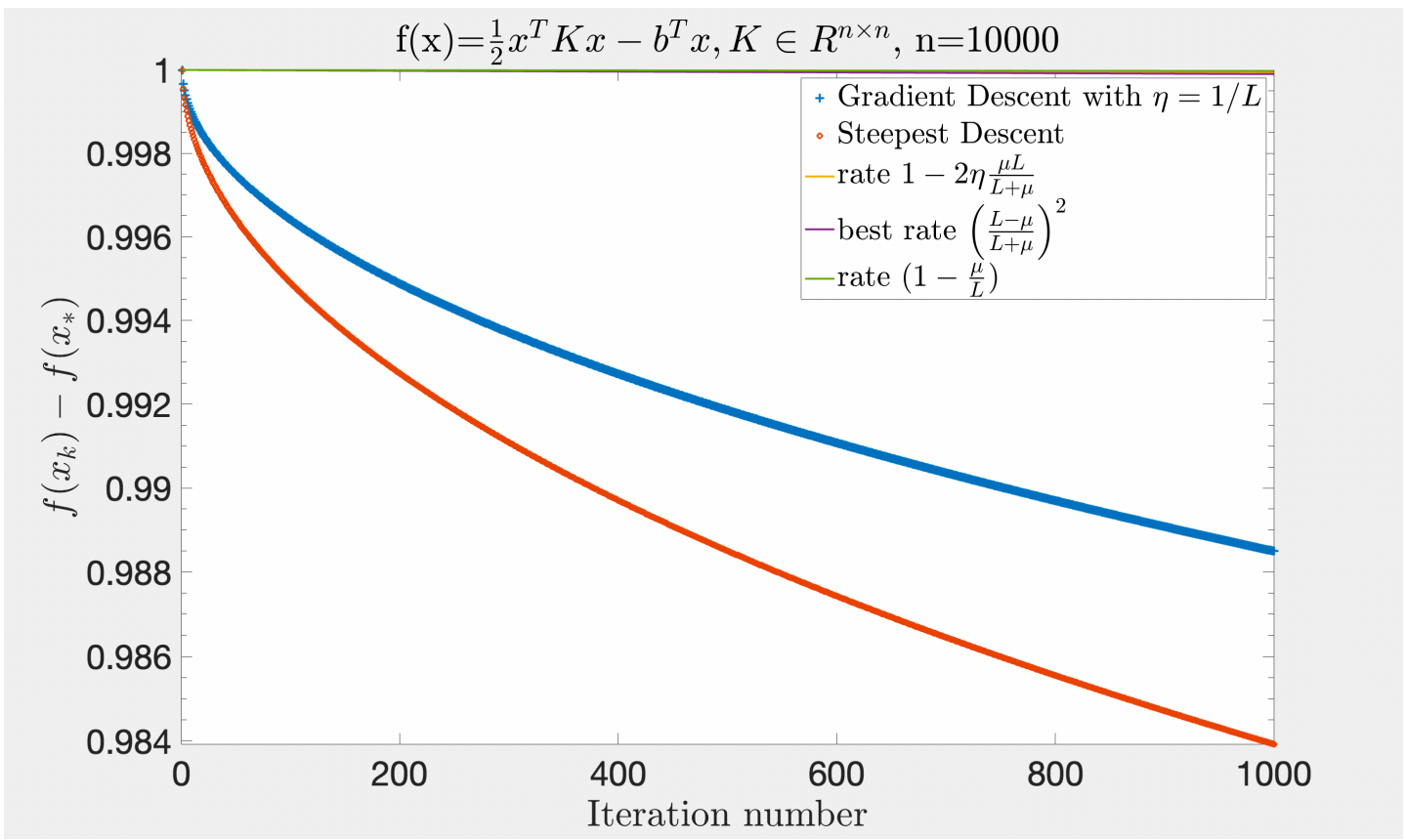
GD with $\eta \leq \frac{2}{L+\mu}$	$C = 1 - 2\eta\frac{\mu L}{L+\mu}$
$\eta = \frac{2}{L+\mu}$	$C = \left(\frac{L-\mu}{L+\mu}\right)^2$
Steepest Descent	$C = 1 - \frac{\mu}{L}$
Steepest Descent for quadratics	$C = \left(\frac{L-\mu}{L+\mu}\right)^2$

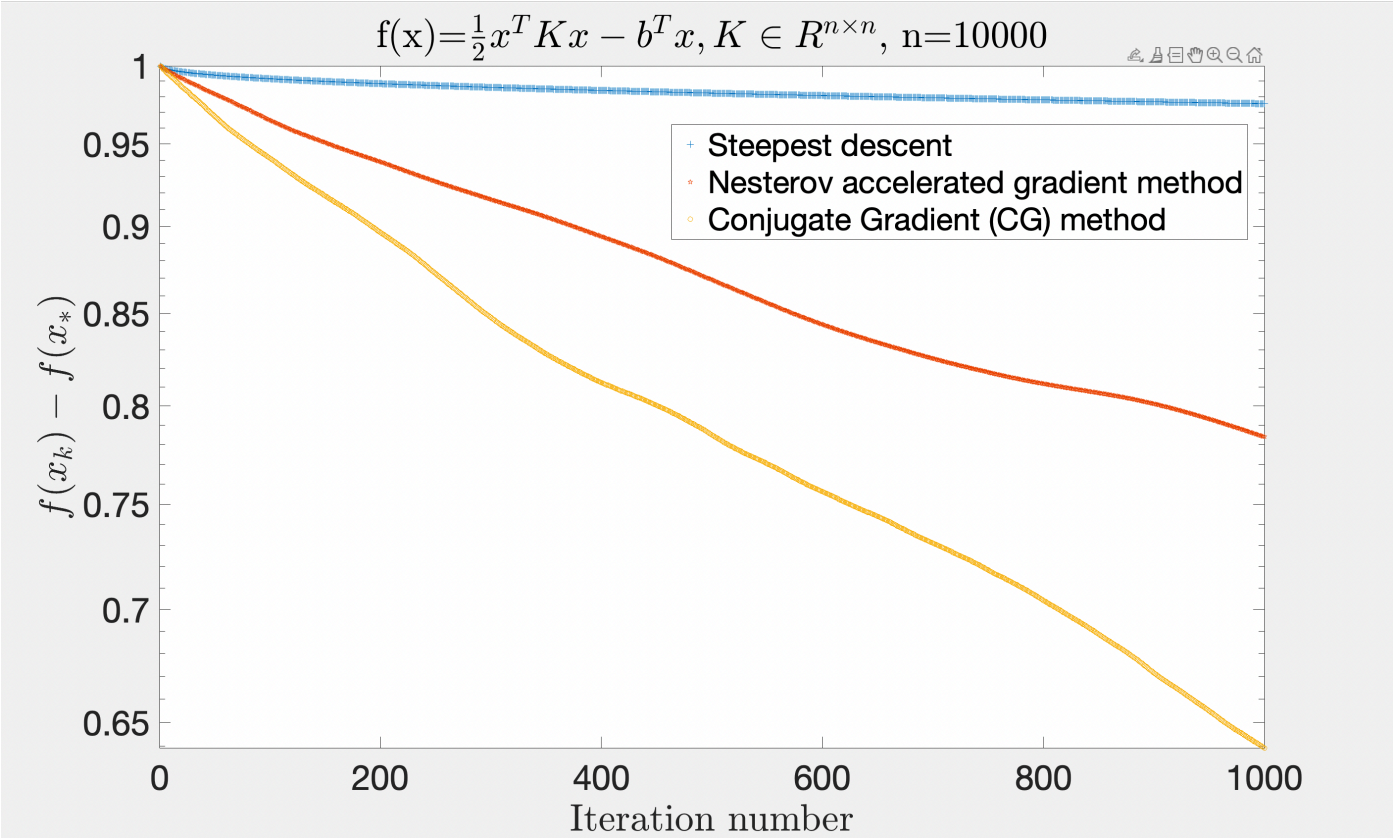



f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=10
Gradient Descent with $\eta = 2/(L+\mu)$
Steepest Descent
rate $1 - 2\eta\frac{\mu L}{L+\mu}$
best rate $\left(\frac{L-\mu}{L+\mu}\right)^2$
rate $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
10^0
10^{-5}
10^{-10}
10^{-15}
0
100
200
300
400
500
600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


f(x)=$\frac{1}{2}x^TKx - b^Tx, K \in R^{n\times n}$, n=100
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 $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
10^0
10^{-1}
0
200
400
600
800
1000
Iteration number


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 $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
Iteration number
1
0.95
0.9
0.85
0.8
0
200
400
600
800
1000


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 $(1 - \frac{\mu}{L})$
$f(x_k) - f(x_*)$
Iteration number
1
0.998
0.996
0.994
0.992
0.99
0.988
0.986
0.984
0
200
400
600
800
1000



Nesterov accelerated gradient method

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

← Next Week

The 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* $\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. *Newton's method* $\mathbf{p}_k = -[\nabla^2 f(\mathbf{x}_k)]^{-1} \nabla f(\mathbf{x}_k)$.
3. *Quasi Newton's method* $\mathbf{p}_k = -B_k \nabla f(\mathbf{x}_k)$, where $B_k \approx [\nabla^2 f(\mathbf{x}_k)]^{-1}$.
4. *Conjugate Gradient Method* $\mathbf{p}_k = -\nabla f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$, 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.

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:

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle, \quad c_1 \in (0, 1). \tag{3.1a}$$

The constant c_1 is usually taken as a small number such as 10^{-4}, and (3.1a) is called the *Armijo condition*. To avoid unacceptably small step sizes, the *curvature condition* requires

$$\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k \rangle \geq c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle, \quad c_2 \in (c_1, 1). \tag{3.1b}$$

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 (3.1b) 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 (3.1) with $0 < c_1 < c_2 < 1$ are called the *Wolfe conditions*.

The following are called the *strong Wolfe conditions*.

$$f(\mathbf{x}_k + \eta_k\mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1\eta_k\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle, \quad c_1 \in (0, 1). \tag{3.2a}$$

$$|\langle \nabla f(\mathbf{x}_k + \eta_k\mathbf{p}_k), \mathbf{p}_k \rangle| \leq c_2|\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle|, \quad c_2 \in (c_1, 1). \tag{3.2b}$$

Lemma 3.1. *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 (3.1) and the strong Wolfe conditions (3.2).*

Proof. The line $\ell(\eta) = f(\mathbf{x}_k) + \eta c_1\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle$ has a negative slope with $0 < c_1 < 1$. So the line must intersect with the graph of $\phi(\eta) = f(\mathbf{x}_k + \eta\mathbf{p}_k)$ at least once for $\eta > 0$, because $0 > \ell'(0) > \phi'(0)$, $\ell(0) = \phi(0)$ and $\phi(\eta)$ is bounded below for all η.

Let $\eta_1 > 0$ be the smallest such intersection point. Then

$$f(\mathbf{x}_k + \eta_1\mathbf{p}_k) = f(\mathbf{x}_k) + \eta_1 c_1\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle,$$

and (3.1a) holds for any $\eta \in (0, \eta_1)$ because $\eta_1 > 0$ is the smallest intersection point.

By the Mean Value Theorem on $\phi(\eta) = f(\mathbf{x}_k + \eta\mathbf{p}_k)$, there is $\eta_2 \in (0, \eta_1)$ such that

$$f(\mathbf{x}_k + \eta_1\mathbf{p}_k) - f(\mathbf{x}_k) = \langle \nabla f(\mathbf{x}_k + \eta_2\mathbf{p}_k), \eta_1\mathbf{p}_k \rangle.$$

By the two equations above, we have

$$\langle \nabla f(\mathbf{x}_k + \eta_2\mathbf{p}_k), \mathbf{p}_k \rangle = c_1\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle > c_2\langle \nabla f(\mathbf{x}_k), \mathbf{p}_k \rangle.$$

So η_2 satisfies (3.1b). Since ∇f is continuous, there is a small interval containing η_2, in which η satisfies (3.1b). Notice that the left hand side of the inequality above is negative, thus the strong Wolfe conditions also hold at η_2 and in a small interval containing η_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 3.3 (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 (3.1). Then*

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(\mathbf{x}_k)\|^2 < +\infty.$$

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \le f(\mathbf{x}_k) + c_1 \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle, \quad c_1 \in (0,1) \tag{3.1a}$$

$$\langle \nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k), \mathbf{p}_k\rangle \ge c_2 \langle \nabla f(\mathbf{x}_k), \mathbf{p}_k\rangle, \quad c_2 \in (c_1, 1) \tag{3.1b}$$

Proof. By (3.1b), we have

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

The Lipschitz continuity and Cauchy-Schwarz inequality give

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

Combining the two inequalities, we get

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

Plugging it into (3.1a), we get

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

which can be written as

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

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

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

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. □

Example 3.4. *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 $B_k = [\nabla^2 f(\mathbf{x}_k)]^{-1}$ *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*

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

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 $\|B_k\| \|B_k^{-1}\| \leq L/\mu$, *and* $\|B_k\| \|B_k^{-1}\|$ *is the condition number of the Hessian. With Theorem 3.3, 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 (3.1) 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 3.5. *Recall that* $\|\nabla f(\mathbf{x}_k)\| \to 0$ *may not even imply* $\mathbf{x}_k$ *converges to a critical point, see Example 2.8. Prove that* $\|\nabla f(\mathbf{x}_k)\| \to 0$ *implies* $\mathbf{x}_k$ *converges to the global minimizer under the assumption*

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