Gradient Descent (GD) Method

- GD is the simplest, but the most popular and the most used method
- It solves unconstrained problems, that is

$$\min_{x \in \mathbb{R}^n} \quad f(x), \quad f : \mathbb{R}^n \to \mathbb{R}$$

- It can extended to solve constrained problems: projected GD
- Some accelerated versions are now available: stochastic GD, AccGD, etc
- It is used to train a neural network/deep network
- It is used to solve a linear regression model
- In general, it is used to solve smooth minimization problems

First/Second Order Necessary Condition for Unconstrained Problem

Consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} \quad f(x), \tag{1}$$

- **FONC:** If $x^* \in \mathbb{R}^n$ is a local minimia of (1), then

  $$\nabla f(x^*) = 0$$

  The point x^* is then called a *stationary point*.
  Note: FONC is not sufficient (e.g. $f(x) = x^3$).
  If f is convex and differentiable, x^* is a global minimia iff $\nabla f(x^*) = 0$.

- **SONC:** If x^* is a local minimizer of the problem (1) then

  $$\nabla^2 f(x^*) \succeq 0.$$

Second Order Sufficient Condition (SOSC)

If x^* is a point such that

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0$$

then x^* is a strict local minimizer.

General Algorithmic Strategies

The goal is to find a local minimizer of the problem

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

GD constructs a sequence of points $\{x^{(k)} : \; k \geq 0\}$ through an iterative approach

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}, \quad k \geq 0$$

- $d^{(k)}$ is a vector with $\|d^{(k)}\| = 1$ (direction of descent)
- $\alpha_k > 0$ (stepsize amount to take along the descent direction)

The idea is that each next point is obtained from the previous point by moving some distance along a direction such that

$$f(x^{(k+1)}) < f(x^{(k)}).$$

and $x^{(k)} \to x^*, \; k \to \infty$, where x^* is the stationary point, i.,e., $\nabla f(x^*) = 0$.

We are going to closely study:

1. How to choose the descent direction
2. How to choose the stepsize selection
3. Stopping Criterion for such method
4. Convergence rate
5. Accelerated variant: Nesterov's method

Descent Direction

A descent direction d is a direction which would decrease the function value, i.e. $f(x^{(k+1)}) < f(x^{(k)})$.

Descent Direction

A descent direction d is a direction which would decrease the function value, i.e. $f(x^{(k+1)}) < f(x^{(k)})$.

Let's denote $f_k = f(x^{(k)})$ and $\nabla f_k = \nabla f(x^{(k)})$.

Descent Direction

A descent direction d is a direction which would decrease the function value, i.e. $f(x^{(k+1)}) < f(x^{(k)})$.

Let's denote $f_k = f(x^{(k)})$ and $\nabla f_k = \nabla f(x^{(k)})$.

By Taylor's Theorem, for $\alpha > 0$ we have

$$f(x^{(k)} + \alpha d) = f(x^{(k)}) + \alpha d^T \nabla f_k + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x^{(k)} + td)d, \quad t \in (0, \alpha)$$

Descent Direction

A descent direction d is a direction which would decrease the function value, i.e. $f(x^{(k+1)}) < f(x^{(k)})$.

Let's denote $f_k = f(x^{(k)})$ and $\nabla f_k = \nabla f(x^{(k)})$.

By Taylor's Theorem, for $\alpha > 0$ we have

$$f(x^{(k)} + \alpha d) = f(x^{(k)}) + \alpha d^T \nabla f_k + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x^{(k)} + td)d, \quad t \in (0, \alpha)$$

Rearrange to get

$$\frac{f(x^{(k)} + \alpha d) - f(x^{(k)})}{\alpha} = d^T \nabla f_k + \frac{1}{2}\alpha d^T \nabla^2 f(x^{(k)} + td)d, \quad t \in (0, \alpha)$$

The rate of change of f along the direction d at $x^{(k)}$ is given by

$$\lim_{\alpha \to 0} \frac{f(x^{(k)} + \alpha d) - f(x^{(k)})}{\alpha} = \nabla f_k^T d.$$

The rate of change of f along the direction d at $x^{(k)}$ is given by

$$\lim_{\alpha \to 0} \frac{f(x^{(k)} + \alpha d) - f(x^{(k)})}{\alpha} = \nabla f_k^T d.$$

Therefore to have a negative rate of change, it is necessary to have

$$\nabla f_k{}^T d < 0$$

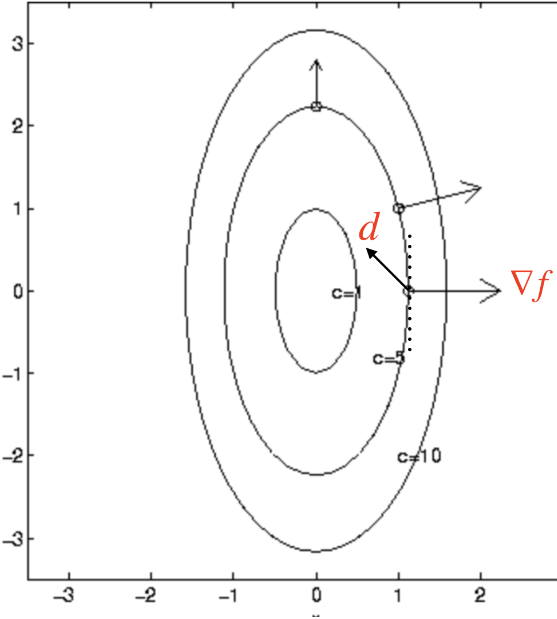
The rate of change of f along the direction d at $x^{(k)}$ is given by

$$\lim_{\alpha \to 0} \frac{f(x^{(k)} + \alpha d) - f(x^{(k)})}{\alpha} = \nabla f_k^T d.$$

Therefore to have a negative rate of change, it is necessary to have

$$\nabla f_k{}^T d < 0$$

This means the angle between ∇f_k and d should be...?


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
c=1
c=5
c=10
d
∇f

What's the "Steepest" Descent Direction

We need to solve the problem

$$\min_d \;\; d^T \nabla f_k, \quad \text{subject to} \quad \|d\| = 1$$

Note that

$$d^T \nabla f_k = \|d\| \|\nabla f_k\| \cos\theta$$

where θ is the angle between d and ∇f_k.

What's the "Steepest" Descent Direction

We need to solve the problem

$$\min_{d} \; d^T \nabla f_k, \quad \text{subject to} \quad \|d\| = 1$$

Note that

$$d^T \nabla f_k = \|d\| \|\nabla f_k\| \cos\theta$$

where θ is the angle between d and ∇f_k.

It is easy to see that the minimizer is attained when $\cos\theta = -1$ and

$$d = -\frac{\nabla f_k}{\|\nabla f_k\|}$$

Method of Steepest Descent

Consider the problem

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

where f is continuously differentiable.

Steepest/gradient descent method:

$$x^{(k+1)} = x^{(k)} - \alpha_k \frac{\nabla f_k}{\|\nabla f_k\|}, \quad k \geq 0$$

In general, you often don't see normalization of gradient, in textbook. However, in practice, sometimes it's better to normalize it.

Illustration of Steepest Descent

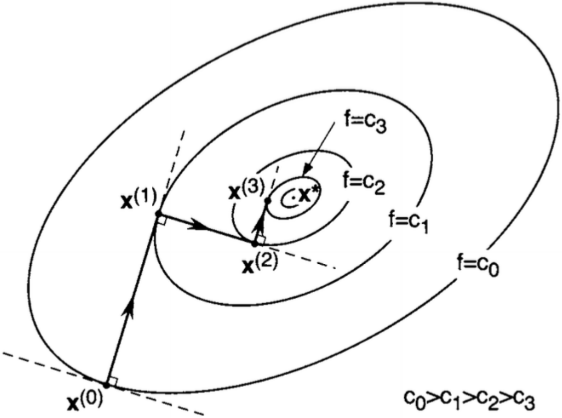
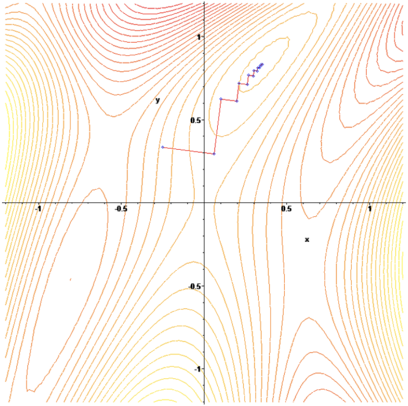
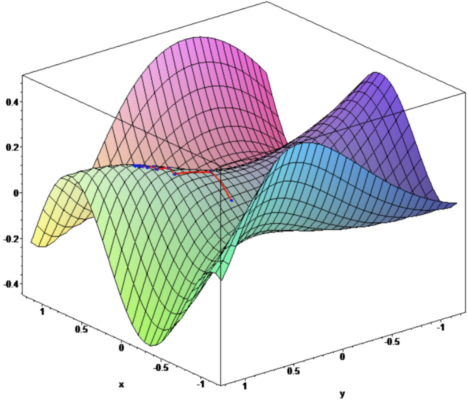


Illustration of Steepest Descent

$$f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$





When to stop the iteration

The first-order necessary condition $\|\nabla f(\boldsymbol{x}^{(k+1)})\| = 0$ is not practical.

Practical conditions:

- gradient condition $\|\nabla f(\boldsymbol{x}^{(k+1)})\| < \epsilon$
- successive objective condition $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})| < \epsilon$ or the relative one

$$\frac{|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|}{|f(\boldsymbol{x}^{(k)})|} < \epsilon$$

- successive point difference $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \epsilon$ or the relative one

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)}\|} < \epsilon$$

- to avoid division by tiny numbers (unstable division), we can replace the denominators by $\max\{1, |f(\boldsymbol{x}^{(k)})|\}$ and $\max\{1, \|\boldsymbol{x}^{(k)}\|\}$, respectively

Stepsize

Small step size:

- Pros: iterations are more likely converge, closely traces max-rate descends
- Cons: need more iterations and thus evaluations of ∇f

Large step size:

- Pros: better use of each $\nabla f(\boldsymbol{x}^{(k)})$, may reduce the total iterations
- Cons: can cause overshooting and zig-zags, too large $\Rightarrow$ diverged iterations

In practice, step sizes are often chosen

- as a fixed value if ∇f is Lipschitz (rate of change is bounded) with the constant known or an upper bound of it known
- by line search
- by a method called Barzilai-Borwein with nonmonotone line search

More Specific Stepsize Choices

- $\alpha_k = \alpha$ fixed small positive, e.g. $\alpha = 10^{-3}$, 10^{-4}, or even smaller

- If f is quadratic, $\alpha_k = \arg\min_{\alpha \geq 0} f\left(x^{(k)} - \alpha \nabla f_k\right)$

- (Backtracking/line search) start with a reasonably big step (e.g. $\alpha = 1$), then gradually reduce it until

$$f(x^{(k)} - \alpha \nabla f_k) < f(x^{(k)}).$$

In practice, you would need to define an inner loop. Fix $t \in (0, 1)$ and $\alpha = 1$, initially. Whenever, $f(x^{(k)} - \alpha \nabla f_k) < f(x^{(k)})$ does not hold, replace $\alpha \leftarrow t\alpha$ to reduce the value of α. Then check the inequality again, if it holds, choose this α to obtain $x^{(k+1)}$. Otherwise, replace $\alpha \leftarrow t\alpha$. Continue this process until the inequality holds.

- (Armijo Line Search) To get a better sufficient decrease in f,

$$f(x^{(k)} - \alpha \nabla f_k) < f(x^{(k)}) + c_1 \alpha \nabla f_k^T d^{(k)}$$

where $c_1 \in (0, 1)$. Note: Similar to backtracking with different inequality.

The reduction of f should be proportional to both the step length and the directional derivative $\nabla f_k^T d^{(k)}$.

Stepsize for Convex Quadratic

For quadratic functions we can find optimal stepsize. Let consider a general convex quadratic function

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

where Q is positive definite matrix. For some $x^{(k)}$, $\nabla f_k = Qx^{(k)} + c$, and $(k+1)$th iteration of the steepest descent is

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f_k$$

Stepsize for Convex Quadratic

For quadratic functions we can find optimal stepsize. Let consider a general convex quadratic function

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

where Q is positive definite matrix. For some $x^{(k)}$, $\nabla f_k = Qx^{(k)} + c$, and $(k+1)$th iteration of the steepest descent is

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f_k$$

To find the optimal stepsize

$$\alpha_k = \arg\min_{\alpha \geq 0} \left\{ f\left(x^{(k)} - \alpha \nabla f_k \right) \right\}.$$

Note that f is differentiable and has an easy structure, so we can find the optimal stepsize.

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2} x^T Q x + c^T x$, then

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2}x^T Qx + c^T x$, then

$$\phi_k(\alpha) = \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q\left(x^{(k)} - \alpha \nabla f_k\right) + c^T\left(x^{(k)} - \alpha \nabla f_k\right)$$

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2}x^T Qx + c^T x$, then

$$\begin{aligned}\phi_k(\alpha) &= \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q\left(x^{(k)} - \alpha \nabla f_k\right) + c^T\left(x^{(k)} - \alpha \nabla f_k\right) \\ &= \alpha^2\left(\frac{1}{2}\nabla f_k^T Q \nabla f_k)\right) - \alpha(\nabla f_k^T \nabla f_k\right) + f(x^{(k)}).\end{aligned}$$

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2}x^T Q x + c^T x$, then

$$\begin{aligned}
\phi_k(\alpha) &= \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q \left(x^{(k)} - \alpha \nabla f_k\right) + c^T \left(x^{(k)} - \alpha \nabla f_k\right) \\
&= \alpha^2 \left(\frac{1}{2} \nabla f_k^T Q \nabla f_k)\right) - \alpha (\nabla f_k^T \nabla f_k\right) + f(x^{(k)}).
\end{aligned}$$

Question 1. Is $\phi_k(\alpha)$ a convex function?

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2}x^T Qx + c^T x$, then

$$\begin{aligned}\phi_k(\alpha) &= \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q\left(x^{(k)} - \alpha \nabla f_k\right) + c^T\left(x^{(k)} - \alpha \nabla f_k\right) \\ &= \alpha^2\left(\frac{1}{2}\nabla f_k^T Q \nabla f_k)\right) - \alpha(\nabla f_k^T \nabla f_k\right) + f(x^{(k)}).\end{aligned}$$

Question 1. Is $\phi_k(\alpha)$ a convex function? Since Q is positive definite, the coefficient of α^2 is positive, so $\phi_k(\alpha)$ is convex quadratic function.

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2}x^T Qx + c^T x$, then

$$\begin{aligned}
\phi_k(\alpha) &= \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q\left(x^{(k)} - \alpha \nabla f_k\right) + c^T\left(x^{(k)} - \alpha \nabla f_k\right) \\
&= \alpha^2\left(\frac{1}{2}\nabla f_k^T Q \nabla f_k)\right) - \alpha(\nabla f_k^T \nabla f_k\right) + f(x^{(k)}).
\end{aligned}$$

Question 1. Is $\phi_k(\alpha)$ a convex function? Since Q is positive definite, the coefficient of α^2 is positive, so $\phi_k(\alpha)$ is convex quadratic function.

Question 2: How can I find the global minimizer?

Let's compute α_k

Define

$$\phi_k(\alpha) := f\left(x^{(k)} - \alpha \nabla f_k\right)$$

Since $f(x) = \frac{1}{2} x^T Q x + c^T x$, then

$$\begin{aligned} \phi_k(\alpha) &= \frac{1}{2}\left(x^{(k)} - \alpha \nabla f_k\right)^T Q \left(x^{(k)} - \alpha \nabla f_k\right) + c^T\left(x^{(k)} - \alpha \nabla f_k\right) \\ &= \alpha^2 \left(\frac{1}{2} \nabla f_k^T Q \nabla f_k)\right) - \alpha(\nabla f_k^T \nabla f_k\right) + f(x^{(k)}). \end{aligned}$$

Question 1. Is $\phi_k(\alpha)$ a convex function? Since Q is positive definite, the coefficient of α^2 is positive, so $\phi_k(\alpha)$ is convex quadratic function.

Question 2: How can I find the global minimizer?

$$\phi'(\alpha) = 0 \qquad \rightarrow \qquad \alpha_k = \alpha = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}.$$

Gradient Descent method for Convex Quadratic Functions

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

The iterates of gradient descent method is given by

$$x^{(k+1)} = x^{(k)} - \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f(x^{(k)})$$

where

$$\nabla f_k = Qx^{(k)} + c$$

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2}x^T Qx + c^T x$, what is Q and c?

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2}x^T Qx + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2}x^T Qx + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

$$Q = 2I_n, \quad c = 0, \quad \nabla f(x) = 2x$$

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2} x^T Q x + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

$$Q = 2I_n, \quad c = 0, \quad \nabla f(x) = 2x$$

Question 3. What's the global minimizer?

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2}x^T Qx + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

$$Q = 2I_n, \quad c = 0, \quad \nabla f(x) = 2x$$

Question 3. What's the global minimizer?

$$\nabla f(x) = 2x = 0 \quad \rightarrow \quad x^* = 0$$

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2}x^T Qx + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

$$Q = 2I_n, \quad c = 0, \quad \nabla f(x) = 2x$$

Question 3. What's the global minimizer?

$$\nabla f(x) = 2x = 0 \quad \rightarrow \quad x^* = 0$$

Recall the steepest descent method

$$x^{(k+1)} = x^{(k)} - \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f(x^{(k)})$$

Example 1

Let's consider the following problem

$$\min f(x) = x^T x$$

Question 1. Comparing $x^T x$ with $\frac{1}{2} x^T Q x + c^T x$, what is Q and c?
Question 2. What's the $\nabla f(x)$?

$$Q = 2I_n, \quad c = 0, \quad \nabla f(x) = 2x$$

Question 3. What's the global minimizer?

$$\nabla f(x) = 2x = 0 \quad \rightarrow \quad x^* = 0$$

Recall the steepest descent method

$$x^{(k+1)} = x^{(k)} - \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f(x^{(k)})$$

Thus for $k = 0$ we obtain

$$x^{(1)} = x^{(0)} - \frac{4(x^{(0)})^T x^{(0)}}{8(x^{(0)})^T x^{(0)}} 2x^{(0)} = x^{(0)} - x^{(0)} = 0.$$

So the global minimizer is obtained in only one step.

Convergence Rate of GD to Convex Quadratic Functions

Theorem

Suppose that

$$\min_x f(x) = \frac{1}{2}x^T Qx + c^T x, \quad Q = Q^T, \quad Q \succeq 0$$

The steepest descent gives the following rate for the function error

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right)[f(x^k) - f(x^*)]$$

where $\lambda_{\min}(Q)$ *and* $\lambda_{\max}(Q)$ *are the smallest and largest eigenvalues of* Q *respectively. Moreover, since* $Qx^* = -c$*, to quantify the rate of convergence we introduce the weighted norm* $\|x\|_Q^2 = x^T Qx$*. Then* $\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*)$*. The rate of convergence in terms of sequence error is*

$$\|x^{k+1} - x^*\|_Q^2 \leq \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right)\|x^k - x^*\|_Q^2.$$

The global minimizer is

$$x^* = -Q^{-1}c,$$

and

$$f(x^*) = -\frac{1}{2}{x^*}^T Q x^*.$$

The global minimizer is

$$x^* = -Q^{-1}c,$$

and

$$f(x^*) = -\frac{1}{2}{x^*}^T Q x^*.$$

For simplicity we consider

$$q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

The global minimizer is

$$x^* = -Q^{-1}c,$$

and

$$f(x^*) = -\frac{1}{2}{x^*}^T Q x^*.$$

For simplicity we consider

$$q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

Note that

$$q(x) = f(x) + \frac{1}{2}{x^*}^T Q x^*,$$

so the two functions differ only by a constant. (Check!)

The global minimizer is

$$x^* = -Q^{-1}c,$$

and

$$f(x^*) = -\frac{1}{2}{x^*}^T Q x^*.$$

For simplicity we consider

$$q(x) = \frac{1}{2}(x - x^*)^T Q (x - x^*)$$

Note that

$$q(x) = f(x) + \frac{1}{2}{x^*}^T Q x^*,$$

so the two functions differ only by a constant. (Check!)

Note that $x^* = -Q^{-1}c$ minimizes both $f(x)$ and $q(x)$ but $q(x^*) = 0$.

Note that

$$f(x^k) - f(x^*) = q(x^k) - \frac{1}{2}{x^*}^T Q x^* - f(x^*) = q(x^{(k)})$$

Note that

$$f(x^k) - f(x^*) = q(x^k) - \frac{1}{2}{x^*}^T Q x^* - f(x^*) = q(x^{(k)})$$

Thus instead of showing

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right)[f(x^k) - f(x^*)]$$

we prove

$$q(x^{k+1}) \leq \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right) q(x^k)$$

Gradient descent for

$$\min_x q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

is then given by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla q_k, \qquad \alpha_k = \frac{\nabla q_k^T \nabla q_k}{\nabla q_k Q \nabla q_k}$$

Gradient descent for

$$\min_x q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

is then given by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla q_k, \qquad \alpha_k = \frac{\nabla q_k^T \nabla q_k}{\nabla q_k Q \nabla q_k}$$

Note that $I = QQ^{-1} = Q^T Q^{-1}$, as Q is symmetric $Q = Q^T$

$$q(x^k) = (x^k - x^*)^T Q(x^k - x^*) \quad = \quad (x^k - x^*)^T Q^T Q^{-1} Q(x^k - x^*)$$

Gradient descent for

$$\min_x q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

is then given by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla q_k, \qquad \alpha_k = \frac{\nabla q_k^T \nabla q_k}{\nabla q_k Q \nabla q_k}$$

Note that $I = QQ^{-1} = Q^T Q^{-1}$, as Q is symmetric $Q = Q^T$

$$\begin{aligned} q(x^k) = (x^k - x^*)^T Q(x^k - x^*) &= (x^k - x^*)^T Q^T Q^{-1} Q(x^k - x^*) \\ &= (Q(x^k - x^*))^T Q^{-1}(Q(x^k - x^*)) \end{aligned}$$

Gradient descent for

$$\min_x q(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$

is then given by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla q_k, \qquad \alpha_k = \frac{\nabla q_k^T \nabla q_k}{\nabla q_k Q \nabla q_k}$$

Note that $I = QQ^{-1} = Q^T Q^{-1}$, as Q is symmetric $Q = Q^T$

$$\begin{aligned} q(x^k) = (x^k - x^*)^T Q(x^k - x^*) &= (x^k - x^*)^T Q^T Q^{-1} Q(x^k - x^*) \\ &= (Q(x^k - x^*))^T Q^{-1}(Q(x^k - x^*)) \\ &= \nabla q_k^T Q^{-1} \nabla q_k \end{aligned}$$

$$q(x^{k+1}) \quad = \quad \frac{1}{2}\left(x^k - \alpha_k \nabla q_k - x^*\right)^T Q\left(x^k - \alpha_k \nabla q_k - x^*\right)$$

$$
\begin{aligned}
q(x^{k+1}) &= \frac{1}{2}\Big(x^k - \alpha_k \nabla q_k - x^*\Big)^T Q \Big(x^k - \alpha_k \nabla q_k - x^*\Big) \\
&= \frac{1}{2}\Big((x^k - x^*) - \alpha_k \nabla q_k\Big)^T Q \Big((x^k - x^*) - \alpha_k \nabla q_k\Big)
\end{aligned}
$$

$$
\begin{aligned}
q(x^{k+1}) &= \frac{1}{2}\Big(x^k - \alpha_k \nabla q_k - x^*\Big)^T Q\Big(x^k - \alpha_k \nabla q_k - x^*\Big) \\
&= \frac{1}{2}\Big((x^k - x^*) - \alpha_k \nabla q_k\Big)^T Q\Big((x^k - x^*) - \alpha_k \nabla q_k\Big) \\
&= q(x^k) - \alpha_k \nabla q_k^T Q(x^k - x^*) + \frac{1}{2}\alpha_k^2 \nabla q_k^T Q \nabla q_k
\end{aligned}
$$

$$
\begin{aligned}
q(x^{k+1}) &= \frac{1}{2}\Big(x^k - \alpha_k \nabla q_k - x^*\Big)^T Q \Big(x^k - \alpha_k \nabla q_k - x^*\Big) \\
&= \frac{1}{2}\Big((x^k - x^*) - \alpha_k \nabla q_k\Big)^T Q \Big((x^k - x^*) - \alpha_k \nabla q_k\Big) \\
&= q(x^k) - \alpha_k \nabla q_k^T Q (x^k - x^*) + \frac{1}{2}\alpha_k^2 \nabla q_k^T Q \nabla q_k \\
&= q(x^k)\Big(1 - \frac{\alpha_k \nabla q_k^T Q (x^k - x^*)}{q(x^k)} + \frac{\frac{1}{2}\alpha_k^2 \nabla q_k^T Q \nabla q_k}{q(x^k)}\Big)
\end{aligned}
$$

Exploit the followings

- $\alpha_k = \frac{\nabla q_k^T \nabla q_k}{\nabla q_k Q \nabla q_k}$

- $Q(x^k - x^*) = \nabla q(x^k)$

- $q(x^k) = \nabla q_k^T Q^{-1} \nabla q_k$

to get

$$q(x^{k+1}) = q(x^k)\left(1 - \frac{\|\nabla q_k\|^4}{(\nabla q_k^T Q \nabla q_k)(\nabla q_k^T Q^{-1} \nabla q_k)}\right)$$

From Rayleigh's inequality,

$$\nabla q_k^T Q \nabla q_k \leq \lambda_{\max}(Q) \|\nabla q_k\|^2$$

$$\nabla q_k^T Q^{-1} \nabla q_k \leq \lambda_{\max}(Q^{-1}) \|\nabla q_k\|^2 \leq (\lambda_{\min}(Q))^{-1} \|\nabla q_k\|^2$$

Therefore,

$$\begin{aligned} q(x^{k+1}) &\leq q(x^k)\Big(1 - \frac{\|\nabla q_k\|^4}{(\nabla q_k^T Q \nabla q_k)(\nabla q_k^T Q^{-1} \nabla q_k)}\Big) \\ &\leq q(x^k)\Big(1 - \frac{\|\nabla q_k\|^4}{\lambda_{\max}(Q)(\lambda_{\min}(Q))^{-1}\|\nabla q_k\|^4}\Big) \\ &= q(x^k)\Big(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\Big) \end{aligned}$$

This terminates the proof.

Moreover...

$$q(x^{k+1}) \leq q(x^k)\Big(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\Big)$$

leads to having

$$q(x^{k+1}) \leq q(x^0)\Big(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\Big)^{k+1}$$

Since $\lambda_{\min}(Q) \leq \lambda_{\max}(Q)$, then

$$1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \leq 1$$

then steepest descent method globally converges.

- If $\lambda_{\min}(Q) = \lambda_{\max}(Q)$, the steepest descent converges in one iterate. For example, $f(x) = x_1^2 + x_2^2$.

Moreover...

$$q(x^{k+1}) \leq q(x^k)\left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right)$$

leads to having

$$q(x^{k+1}) \leq q(x^0)\left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}\right)^{k+1}$$

Since $\lambda_{\min}(Q) \leq \lambda_{\max}(Q)$, then

$$1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \leq 1$$

then steepest descent method globally converges.

- If $\lambda_{\min}(Q) = \lambda_{\max}(Q)$, the steepest descent converges in one iterate. For example, $f(x) = x_1^2 + x_2^2$.
- If $\lambda_{\max}(Q)$ is much larger than $\lambda_{\min}(Q)$, then

$$1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \approx 1$$

then convergence can be extremely slow.

Convergence Rate of GD with line search

Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by the steepest descent method with exact line searches converge to a point x^* at which the Hessian $\nabla^2 f(x^*)$ is positive definite. Let r be any scalar satisfying

$$r \in \Big(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\Big),$$

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(x^*)$. Then for all k sufficiently large, we have

$$f(x^{k+1}) - f(x^*) \leq r^2(f(x^k) - f(x^*))$$

This theorem shows that the steepest descent can have an unacceptably slow rate of convergence, even when the Hessian is reasonably well conditioned. For example, if $\kappa(Q) = 800$, and $f(x^1) = 1$ and $f(x^*) = 0$.

Accelerated Steepest Descent by Nesterov

$$x^k = y^k - \alpha \nabla f(y^k)$$

$$\delta^{k+1} = \frac{1+\sqrt{1+4(\delta^k)^2}}{2}$$

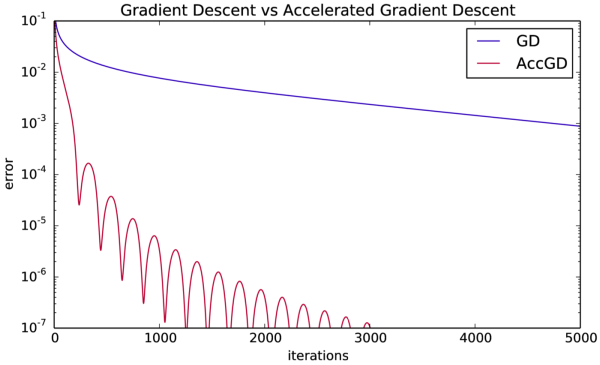
$$y^{k+1} = x^k + \frac{\delta^k - 1}{\delta^{k+1}}(x^k - x^{k-1}) \quad \text{momentum step}$$

The method is initialized with $x^0 = y^1$ and $\delta^1 = 1$, and the the first iteration has index $k = 1$.

Theorem

Let f be a convex and β-smooth function, then Nesterov's Accelerated Gradient Descent satisfies

$$f(y^k) - f(x^*) \leq \frac{2\beta\|x^1 - x^*\|^2}{k^2}$$



Source: http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html