Notes on Convex Optimization
Gradient Descent

Chunpai Wang
cwang25@albany.edu

Spring 2016

1 Objective Function

Consider unconstrained, smooth convex optimization

$$\min \ f(x) \tag{1}$$

i.e., f is convex and differentiable with $dom(f) = \mathbb{R}^n$. Denote the optimal criterion value by $p^* = \min f(x)$, and the optimal solution by x^*.

2 Intuition and Interpretation

Gradient Descent Method: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k+1)} = x^{(k)} - t * \nabla f(x^{(k)}), \ k = 1, 2, 3, ... \tag{2}$$

uitl converge, where $t > 0$ is the step length. Here we may have few questions:

- Why does this update guarantee the descent ?
- What assumptions do we require ?
- How can we decide the value of t ?

Recall, first order Taylor expansion:

$$f(y) = f(x) + \nabla f(x)^T (y - x) + o(\|y - x\|_2^2) \tag{3}$$

and second order Taylor expansion:

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|_2^2) \tag{4}$$

You can refer here for little-o notation.

Most intuitively way to validate the descent is by comparing the value of $f(x^{(k)})$ and $f(x^{(k-1)})$.

$$\begin{aligned}
f(x^{(k+1)}) &= f(x^{(k)}) + \nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}) + o(\|x^{(k+1)} - x^{(k)}\|_2) && (5)\\
&= f(x^{(k)}) + \nabla f(x^{(k)})(x^{(k)} - t\nabla f(x^{(k)}) - x^{(k)}) + o(\|x^{(k)} - t\nabla f(x^{(k)} - x^{(k)}\|_2) && (6)\\
&= f(x^{(k)}) + \nabla f(x^{(k)})(-t\nabla f(x^{(k)})) + o(\| - t\nabla f(x^{(k)}\|_2) && (7)\\
&= f(x^{(k)}) - t\nabla \|f(x^{(k)})\|_2^2 + o(\| - t\nabla f(x^{(k)}\|_2) && (8)\\
&\leq f(x^{(k)}) && (9)
\end{aligned}$$

Since norm is non-negative, we can conclude $f(x^{(k+1)}) \leq f(x^{(k)})$ with the gradient descent updating rule (2).

Another interpretation is, gradient descent updating rule is the closed form solution to minimize the approximation of second order Taylor expansion. First, we can find quadratic approximation of second order Taylor expansion by replacing the usual Hessian $\nabla^2 f(x)$ with $\frac{1}{t}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence, at each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}||y - x||_2^2 = g(y) \tag{10}$$

We can see, there is a linear approximation term: $f(x) + \nabla f(x)^T(y-x)$, and proximity penalty term to x, with weight $1/(2t)$ that is $\frac{1}{2t}||y-x||_2^2$. Then, we can try to minimize the approximation $g(y)$ by taking derivative w.r.t. y and setting to zero, we have

$$\begin{aligned} \nabla g(y) &= 0 \\ \nabla f(x) + \frac{1}{t}(y-x) &= 0 \\ y &= x - t\nabla f(x) \end{aligned} \tag{11}$$

Thus, we can interpret the gradient descent as a method to minimize the quadratic approximation of the function, which passes through point $(x, f(x))$. In addition, we can realize the importance of step length t in the figure below.

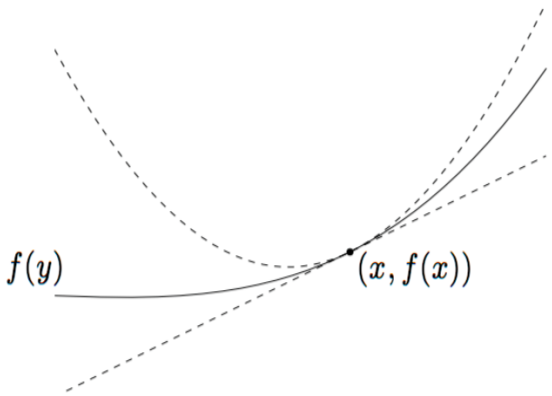

Figure 1: We can find that the value of t actually determine whether we solve the quadratic upper bound of f or lower bound of f, and it is also called the learning rate.

3 Line Search

How to choose a reasonable step size is worth to grasp, since if step size is too large, the descent will never converge, and if the step size is too small, it will take forever to converge. Usually, there are 3 strategies to choose the step size.

- fixed constant t
- exact line search: minimize $f(x - t\nabla f(x))$ over t, but this is usually very expensive to find.
- backtracking line search

3.1 Backtracking Line Search

- First fix parameters $0 < \beta < 1$ and $0 < \alpha < 1/2$
- At each iteration, start with $t = 1$, and while

$$f(x - t\nabla f(x)) \leq f(x) - \alpha t||\nabla f(x)||_2^2 \tag{12}$$

  shrink $t = \beta t$, which means check if the objective decrease at least $\alpha t||\nabla f(x)||_2^2$ amount. If not, shrink $t = \beta t$ to make it converge fast, and check it again. Otherwise, perform gradient descent update

$$x^+ = x - t\nabla f(x)$$

We can visualize the backtracking line search as following:

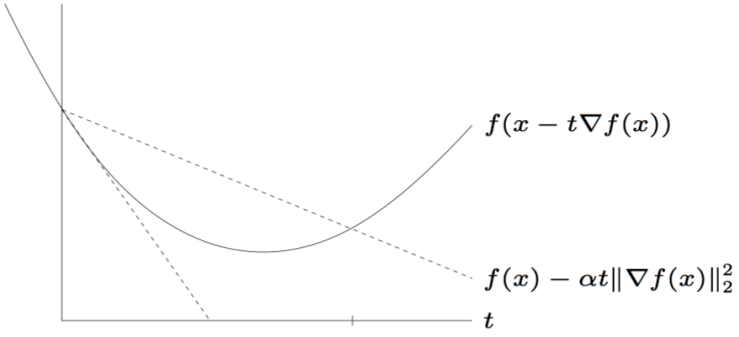

Figure 2: Backtracking Line Search Interpretation

4 Convex Function

In order to prove a function is convex, we need to first show the domain of f is convex. Then, we can use 5 ways to prove it:

1. Jensen's inequality
2. First order condition
3. Second order condition
4. Monotonicity of gradient
5. Midpoint convex
6. Restriction to a line

Convex Function: f is convex if dom(f) is a convex set and Jensen's inequality holds:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in dom(f), \theta \in [0,1] \tag{13}$$

First-order condition: for (smooth or continuously) differentiable f, Jensen's inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) \quad \forall x, y \in dom(f) \tag{14}$$

Second-order condition: for twice differentiable f, Jensen's inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in dom(f) \quad \text{(Positive Semi-definite)} \tag{15}$$

Monotonicity of gradient: differentiable f is convex if and only if dom(f) is convex and

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq 0 \quad \forall x, y \in dom(f) \tag{16}$$

which means if x greater than y, then $\nabla f(x)$ is also greater than $\nabla f(y)$; i.e., $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone mapping.

Proof. ($\Rightarrow$) if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^T(y-x), \quad f(x) \geq f(y) + \nabla f(y)^T(x-y)$$

combining the inequalities gives $\nabla f(x) - \nabla f(y))^T(x-y) \geq 0$.

($\Leftarrow$) if ∇f is monotone, then $g'(t) \geq g'(0)$ for all $t \geq 0$ and $t \in dom(g)$, where

$$g(t) = f(x + t(y-x)), \quad g'(t) = \nabla f(x + t(y-x))^T(y-x)$$

Hence,

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)dt \quad \geq \quad g(0) + g^{'}(0) = f(x) + \nabla f(x)^T(y-x)$$

□

Midpoint Convex: A function $f : I \to \mathbb{R}$ on interval I called mid-point convex if

$$f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2} \quad \forall x, y \in I \tag{17}$$

If f is continuous midpoint convex, then f is convex.

Restriction to a line: $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for any $x \in dom(f)$ and $v \in \mathbb{R}^n$, the function $g : \mathbb{R} \to \mathbb{R}$ where $g(t) = f(x + tv)$ is convex in t with $dom(g) = \{t | x + tv \in dom(f)\}$.

Proof. ($\Rightarrow$) Assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $t_1, t_2 \in dom(g)$, and $\theta \in [0, 1]$, we will show that g is convex.

$$\begin{aligned} g(\theta t_1 + (1-\theta)t_2) &= f(x + (\theta t_1 + (1-\theta)t_2)v) \\ &= f(\theta(x + t_1 v) + (1-\theta)(x + t_2 v)) \\ &\leq \theta f(x + t_1 v) + (1-\theta) f(x + t_2 v) \\ &= \theta g(t_1) + (1-\theta) g(t_2) \end{aligned}$$

($\Leftarrow$) □

5 Quadratic Upper Bound on Convex Functions

Lipschitz Continuous Gradient: gradient of f is Lipschitz continuous with parameter $L \geq 0$ if

$$||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2 \quad \forall x, y \in dom(f) \tag{18}$$

- note that the definition does not assume convexity of f
- we will see that for convex f with $dom(f) = \mathbb{R}^n$, this is equivalent to

$$\frac{L}{2} x^T x - f(x) \text{ is convex}$$

(i.e. if f is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all x)

Quadratic Upper Bound: suppose ∇f is Lipschitz continuous with parameter L and $dom(f)$ is convex

- then $g(x) = \frac{L}{2} x^T x - f(x)$, with $dom(g) = dom(f)$, is convex.
- convexity of g is equivalent to a quadratic upper bound on f:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}||y - x||_2^2 \quad \forall x, y \in dom(f) \tag{19}$$

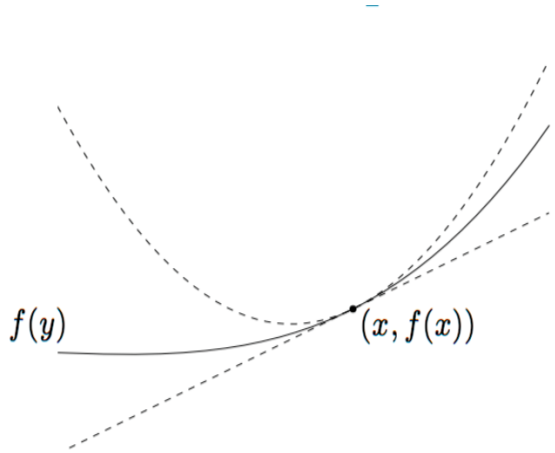

Figure 3: Visualization of Quadratic Upper Bound

Proof. Assume ∇f is Lipschitz continuous with parameter L and $dom(f)$ is convex. And assume $g(x) = \frac{L}{2} x^T x - f(x)$, with $dom(g) = dom(f)$, then g is differentiable and $\nabla g(x) = Lx - \nabla f(x)$.

$$||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2 \quad \forall x, y \in dom(f)$$

$$||\nabla f(x) - \nabla f(y)||_2 * ||x-y||_2 \leq L||x-y||_2 * ||x-y||_2$$

According to Cauchy-Schwarz inequality:

$$|(\nabla f(x) - \nabla f(y))^T (x-y)| \leq ||\nabla f(x) - \nabla f(y)||_2 * ||x-y||_2$$

and since there is no assumption about convexity of f, therefore $(\nabla f(x) - \nabla f(y))^T (x-y)$ could be less equal to 0. However, it still holds for:

$$\begin{aligned}
&\Rightarrow (\nabla f(x) - \nabla f(y))^T (x-y) \leq L||x-y||_2^2 \\
&\Leftrightarrow L||x-y||_2^2 - (\nabla f(x) - \nabla f(y))^T (x-y) \geq 0 \\
&\Leftrightarrow L(x-y)^T(x-y) - (\nabla f(x) - \nabla f(y))^T (x-y) \geq 0 \\
&\Leftrightarrow ((L(x-y) - (\nabla f(x) - \nabla f(y)))^T (x-y) \geq 0 \\
&\Leftrightarrow ((Lx - \nabla f(x)) - (Ly - \nabla f(y))^T (x-y) \geq 0 \\
&(\nabla g(x) - \nabla g(y))^T (x-y) \geq 0 \quad \forall x, y \in dom(f)
\end{aligned}$$

Therefore, according to Monotonicity of Gradient theorem and $dom(g) = dom(f)$ is convex, g is convex function. Since the convexity of g, we can use equation(8) and get

$$g(y) \geq g(x) + \nabla g(x)^T (y-x)$$

$$\begin{aligned}
\frac{L}{2} y^T y - f(y) &\geq \frac{L}{2} x^T x - f(x) + (Lx - \nabla f(x))^T (y-x) \\
-f(y) &\geq \frac{L}{2} x^T x - \frac{L}{2} y^T y - f(x) + (Lx - \nabla f(x))^T (y-x) \\
f(y) &\leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} y^T y - \frac{L}{2} x^T x - Lx^T (y-x) \\
f(y) &\leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} y^T y - \frac{L}{2} x^T x - Lx^T y + Lx^T x) \\
f(y) &\leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} ||y-x||_2^2
\end{aligned}$$

□

Consequence of Quadratic Upper Bound: after derived the quadratic upper bound(dashed-curve in figure2), we are trying to find the optimal point to minimize the dashed-curve. Denote dashed-curve as function g and solid-curve as function f, and $(x, f(x))$ is current point. If $dom(f) = \mathbb{R}^n$ and f has a minimizer x^*, then

$$\frac{1}{2L} ||\nabla f(x)||_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2} ||x - x^*||_2^2 \tag{20}$$

Any interpretation about this inequalities ?
Right-hand inequality: follows from quadratic upper bound at $x = x^*$ and $\nabla f(x^*) = 0$. It is like trying to find the upper bound of f at optimal point x^*.
Left-hand inequality: follows by minimizing quadratic upper bound

$$\begin{aligned}
f(x^*) \leq \ &\mathbf{inf}_{y \in dom(f)} \ \left(f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} ||y-x||_2^2 \right) \\
&= f(x) - \frac{1}{2L} ||\nabla f(x)||_2^2
\end{aligned}$$

minimizer of upper bound is $y = x - \frac{1}{L} \nabla f(x)$, just take the derivative of upper bound with respect to y.

Co-coercivity of Gradient: if f is convex with $dom(f) \in \mathbb{R}^n$ and $\frac{L}{2} x^T x - f(x)$ is convex then

$$\big(\nabla f(x) - \nabla f(y)\big)^T (x-y) \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2^2 \quad \forall x, y \tag{21}$$

this property is known as co-coercivity of ∇f (with parameter $1/L$)

- co-coercivity implies Lipschitz continuity of ∇f (by Chauchy-Schwarz)

- hence, for differentiable convex f with $dom(f) = \mathbb{R}^n$

$$\begin{aligned}\text{Lipschitz continuity of } \nabla f &\Rightarrow \text{convexity of} (L/2)x^T x - f(x) \\ &\Rightarrow \text{co-coercivity of } \nabla f \\ &\Rightarrow \text{Lipschitz continuity of } \nabla f\end{aligned}$$

therefore, these three properties are equivalent.

Proof. We have already proved the (Lipschitz continuity of $f \Rightarrow$ convexity of $(L/2)x^T x - f(x)$). Now we continue to prove (convexity of $(L/2)x^T x - f(x) \Rightarrow$ co-coercivity of ∇f) part. Here define convex functions f_x, f_y with domain R^n:

$$f_x(z) = f(z) - \nabla f(x)^T z, f_y(z) = f(z) - \nabla f(y)^T z$$

Since $(L/2)z^T z$, f_x and f_y are all convex, the functions $(L/2)z^T z - f_x(z)$ and $(L/2)z^T z - f_y(z)$ are also convex. Take derivative of $f_x(z)$ and $f_y(z)$ with respect to z, we can find that:

- $z = x$ minimize the $f_x(z)$. Hence,

$$\begin{aligned}f(y) - f(x) - \nabla f(x)^T (y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L} ||\nabla f_x(y)||_2^2 \quad \text{(inequality in formula(13))} \\ &= \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2\end{aligned}$$

- similarly, $z = y$ minimize the $f_y(z)$; therefore

$$f(x) - f(y) - \nabla f(y)^T (x - y) \geq \frac{1}{2L} ||\nabla f(y) - \nabla f(x)||_2^2$$

- combine above two inequalities shows co-coercivity.

Now we continue to prove the (co-coercivity of $\nabla f \Rightarrow$ Lipschitz continuity of ∇f). Assume the co-coercivity of ∇f, that is

$$\big(\nabla f(x) - \nabla f(y)\big)^T (x - y) \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2^2 \quad \forall x, y \in dom(f)$$

According to Chauchy-Schwarz inequality,

$$||\nabla f(x) - \nabla f(y)||_2 ||x - y||_2 \geq \big(\nabla f(x) - \nabla f(y)\big)^T (x - y) \geq \frac{1}{L} ||\nabla f(x) - \nabla f(y)||_2^2 \quad \forall x, y \in dom(f)$$

Therefore,

$$||\nabla f(x) - \nabla f(y)||_2 \leq L ||x - y||_2 \quad \forall x, y \in dom(f)$$

, which is the Lipschitz continuity of ∇f □

6 Quadratic Lower Bound on Strongly Convex Functions

Strongly Convex Function: f is strongly convex with parameter $m > 0$ if

$$g(x) = f(x) - \frac{m}{2} x^T x \quad \text{is convex} \tag{22}$$

That means, strongly convex f satisfies following properties different with pure convex function.

Jensen's inequality: Jensen's inequality for g is

$$g(\theta x + (1 - \theta) y) \leq \theta g(x) + (1 - \theta) g(y)$$

$$\Rightarrow f(\theta x + (1 - \theta) y) \leq \theta f(x) + (1 - \theta) f(y) - \frac{m}{2} \theta (1 - \theta) ||x - y||_2^2 \tag{23}$$

First-order condition: for 1st order condition of convexity of g:

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \quad \forall x, y \in dom(g)$$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} ||y - x||_2^2 \quad \forall x, y \in dom(f) \tag{24}$$

You can see that the curve of f is pretty steep, which is another view of "strong".
Monotonicity: monotonicity of ∇g gives

$$(\nabla g(x) - \nabla g(y))^T (x - y) \geq 0$$

$$\Rightarrow (\nabla f(x) - \nabla f(y))^T (x - y) \geq m||x - y||_2^2 \quad \forall x, y \in dom(f) \tag{25}$$

this is called strong monotonicity (coercivity) of ∇f

Second-order condition: $\nabla^2 f(x) \succeq mI \quad \forall x \in dom(f)$

Epigraph, Level Sets, & Sublevel Sets
Epigraph: the epigraph, denoted $epi(f)$, describes the set of input-output pairs that f can achieve, as well as "anything above"

$$epi(f) := \{(x, t) \mid x \in dom(f), f(x) \leq t\} \tag{26}$$

Level sets: level sets are sets of points that achieve exactly a certain value for f. Precisely, the t-level set of f is defined by

$$L_t(f) := \{x \in dom(f) \mid f(x) = t\} \tag{27}$$

Sub-level Sets: t-sub-level set of f is defined by

$$S_t(f) := \{x \in dom(f) \mid f(x) \leq t\} \tag{28}$$

Notice the difference between definitions of epigraph and sub-level sets.
Quadratic Lower Bound: from 1st order condition of convexity of g (equation(17)):

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}||y - x||_2^2 \quad \forall x, y \in dom(f)$$

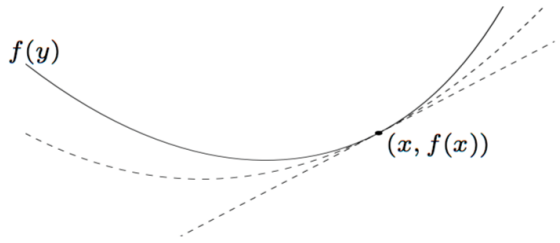

Figure 4: Quadratic Lower Bound of f

- implies sublevel sets of f are bounded. (why?)
- if f is closed (has closed sublevel sets), it has a unique minimizer x^* and

$$\frac{m}{2}||x - x^*||_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}||\nabla f(x)||_2^2 \quad \forall x \in dom(f)$$

Left-hand inequality: follows from quadratic lower bound $x = x^*$. It is trying to find the lower bound of f at optimal point x^*, and $\nabla f(x^*) = 0$

Right-hand inequality: follows by minimizing quadratic lower bound

$$\begin{aligned} f(x^*) &\geq \mathbf{inf}_{y \in dom(f)} \quad \left(f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}||y - x||_2^2\right) \\ &= f(x) + \nabla f(x)^T (x - \frac{1}{m}\nabla f(x) - x) + \frac{m}{2}||x - \frac{1}{m}\nabla f(x) - x||_2^2 \\ &= f(x) - \frac{1}{m}||\nabla f(x)||_2^2 + \frac{1}{2m}||\nabla f(x)||_2^2 \\ &= f(x) - \frac{1}{2m}||\nabla f(x)||_2^2 \end{aligned}$$

minimizer of lower bound is $y = x - 1/m\nabla f(x)$.

Extension of co-coercivity if f is strongly convex and ∇f is Lipschitz continuous, then

- $g(x) = f(x) - \frac{m}{2}x^T x \quad$ is convex, and ∇g is Lipschitz continuous with parameter $L - m$.

- and co-coercivity of g gives:

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{mL}{m+L}||x-y||_2^2 + \frac{1}{m+L}||\nabla f(x) - \nabla f(y)||_2^2 \quad \forall x, y \in dom(f) \tag{29}$$

Proof. Assume f is strongly convex, that is

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)||x-y||_2^2$$

$$\begin{aligned}
g(\theta x + (1-\theta)y) &= f(\theta x + (1-\theta)y) - \frac{m}{2}||\theta x + (1-\theta)y||_2^2 \\
&\leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)||x-y||_2^2 - \frac{m}{2}||\theta x + (1-\theta)y||_2^2 \\
&= \theta f(x) + (1-\theta)f(y) - \big(\theta\frac{m}{2}||x||_2^2 + (1-\theta)\frac{m}{2}||y||_2^2\big) \\
&= \theta f(x) - \theta\frac{m}{2}||x||_2^2 + (1-\theta)f(y) - (1-\theta)\frac{m}{2}||y||_2^2 \\
&= \theta g(x) + (1-\theta)g(y)
\end{aligned}$$

Therefore, $g(x)$ is convex. Now assume ∇f is Lipschitz continuous, that is, existing a paramemter $L > 0$ such that $||\nabla f(x) - \nabla f(y)||_2 \leq L||x-y||_2 \quad \forall x, y \in dom(f)$

$$\begin{aligned}
||\nabla g(x) - \nabla g(y)||_2^2 &= ||\nabla f(x) - mx - \nabla f(y) + my||_2^2 \\
&= ||\nabla f(x) - \nabla f(y) - m(x-y)||_2^2 \\
&= ||\nabla f(x) - \nabla f(y)||_2^2 - 2m(\nabla f(x) - \nabla f(y))^T(x-y) + ||m(x-y)||_2^2 \\
&= ||\nabla f(x) - \nabla f(y)||_2(||\nabla f(x) - \nabla f(y)||_2 - 2m||x-y||_2) + m^2||x-y||_2^2 \\
&\leq L||x-y||_2(L||x-y||_2 - 2m||x-y||_2) + m^2||x-y||_2^2 \\
&= L^2||x-y||_2^2 - 2mL||x-y||_2^2 + m^2||x-y||_2^2 \\
&= (L-m)^2||x-y||_2^2
\end{aligned}$$

hence, $||\nabla g(x) - \nabla g(y)||_2 \leq (L-m)||x-y||_2$, which also means ∇g is Lipschitz continuous with parameter $L-m$.
Since g is convex, ∇g is Lipschitiz continuous $\Leftrightarrow$ co-coercivity of ∇g (according to three equivalent properties). Hence, formula (14) changes to

$$(\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-m}||\nabla g(x) - \nabla g(y)||_2^2$$

$$(\nabla f(x) - \nabla f(y) - m(x-y))^T(x-y) \geq \frac{1}{L-m}||\nabla f(x) - \nabla f(y) - m(x-y)||_2^2$$

$$(1 + \frac{2m}{L-m})(\nabla f(x) - \nabla f(y))^T(x-y) \geq (m + \frac{m^2}{L-m})||x-y||_2^2 + \frac{1}{L-m}||\nabla f(x) - \nabla f(y)||_2^2$$

$$(\frac{L+m}{L-m})(\nabla f(x) - \nabla f(y))^T(x-y) \geq (\frac{mL}{L-m})||x-y||_2^2 + \frac{1}{L-m}||\nabla f(x) - \nabla f(y)||_2^2$$

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq (\frac{mL}{L+m})||x-y||_2^2 + \frac{1}{L+m}||\nabla f(x) - \nabla f(y)||_2^2$$

□

7 Convergence Analysis For Convex Functions

Assumptions:

- f is convex and differentiable with $dom(f) = R^n$
- $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$.
- optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*

For Constant Step Size:
Theorem: Gradient descent with fixed step size $t < 1/L$ satisfies

$$f(x^k) - f^* \leq \frac{||x^{(0)} - x^*||_2^2}{2tk} \tag{30}$$

that is gradient descent has convergence rate $O(1/k)$, i.e., to get $f(x^k) - f^* \leq \epsilon$ we need $O(1/\epsilon)$ iterations.

Proof. ∇f Lipschitz with constant $L \Rightarrow$ quadratic upper bound of f:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}||y - x||_2^2 \quad \forall x, y \in dom(f)$$

Plugging in $y = x - t\nabla f(x)$:

$$f(x - t\nabla f(x)) \leq f(x) - t(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \tag{31}$$

Therefore, if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 1/L$ (that is $Lt < 1 \Rightarrow t(1 - \frac{Lt}{2}) > \frac{t}{2}$)

$$\begin{aligned}
f(x^+) &\leq f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \quad \text{(Note this step, compared with formula(27) in backtracking line search)} \\
&\leq f^* + \nabla f(x)^T(x - x^*) - \frac{t}{2}||\nabla f(x)||_2^2 \quad \text{(Because f(x) - } \nabla f(x)^T(x - x^*) < f^*\text{)} \\
&= f^* + [\frac{1}{2}\nabla f(x)]^T[2(x - x^*)] - [\frac{1}{2}\nabla f(x)]^T[t\nabla f(x)] \\
&= f^* + (\frac{1}{2}\nabla f(x))^T(2x - 2x^* - t\nabla f(x)) \\
&= f^* + \frac{1}{2t}(t\nabla f(x))^T(2x - 2x^* - t\nabla f(x)) \\
&= f^* + \frac{1}{2t}\big(||x - x^*||_2^2 - ||x - x^* - t\nabla f(x)||_2^2\big) \quad \text{(according to } (a-b)(a+b) = a^2 - b^2\text{)} \\
&= f^* + \frac{1}{2t}\big(||x - x^*||_2^2 - ||x^+ - x^*||_2^2\big)
\end{aligned}$$

Take $x = x^{(i-1)}, x^+ = x(i), t_i = t$ and add the bounds for i = 1,...,k:

$$\begin{aligned}
\sum_{i=1}^{k}\big(f(x^{(i)} - f^*\big) &\leq \frac{1}{2t}\sum_{i=1}^{k}\big(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2\big) \\
&= \frac{1}{2t}\sum_{i=1}^{k}\big(||x^{(0)} - x^*||_2^2 - ||x^{(k)} - x^*||_2^2\big) \\
&\leq \frac{1}{2t}||x^{(0)} - x^*||_2^2
\end{aligned}$$

Since $f(x^{(i)})$ is non-increasing,

$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \leq \frac{1}{2kt}||x^{(0)} - x^*||_2^2 \tag{32}$$

Therefore, number of iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\epsilon)$ (WHY??) □

For Backtracking Line Search:
Theorem: Gradient descent with backtracking line search satisfies

$$f(x^{(k)}) - f^* \leq \frac{||x^{(0)} - x^*||_2^2}{2t_{min}k} \tag{33}$$

where $t_{min} = \mathbf{min}\{1, \frac{\beta}{L}\}$. If β is not too small, then we don't lose much compared to fixed step size (β/L vs $1/L$).

Proof. We knew that backtracking line search is to initialize $t_1 > 0$ (usually, $t_1 = 1$); take $t_k := \beta t_k$ until

$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k ||\nabla f(x)||_2^2 \tag{34}$$

Now take $0 < \beta < 1; \alpha = 1/2$ (to simplify proofs),gives

$$\begin{aligned} f(x^{(i)}) &\leq f^* + \frac{1}{2t_i}\big(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2\big) \\ &\leq f^* + \frac{1}{2t_{min}}\big(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2\big) \end{aligned}$$

Since $f(x^{(i)})$ is non-increasing, add the upper bounds to get

$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^*) \leq \frac{1}{2kt_{min}}||x^{(0)} - x^*||_2^2 \tag{35}$$

□

7.1 Convergence Analysis For Strongly Convex Functions

Assumptions:

- f is strongly convex and differentiable with $dom(f) = R^n$
- $\nabla f(x)$ is Lipschitz continuous with parameter $L > 0$.
- optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*

For Constant Step Size:
Theorem: Gradient descent with fixed step size $t \leq 2/(m + L)$ or with backtracking line search satisfies

$$f(x^{(k)}) - f^* \leq c^k \frac{L}{2}||x^{(0)} - x^*||_2^2 \tag{36}$$

with $0 < c < 1$. That is, converge rate with strong convexity is $O(c^k)$, exponentially fast. Therefore, in order to get $f(x^{(k)}) - f^* \leq \epsilon$, it needs $O(log(1/\epsilon))$ iterations.

Proof. if $x^+ = x - t\nabla f(x)$ and $0 < t < \frac{2}{L+m}$:

$$\begin{aligned} ||x^+ - x^*||_2^2 &= ||x - x^* - t\nabla f(x)||_2^2 \\ &= ||x - x^*||_2^2 - 2t\nabla f(x)^T(x - x^*) + t^2||\nabla f(x)||_2^2 \\ &(\text{since}(\nabla f(x) - \nabla f(x^*))^T(x - x^*) = \nabla f(x)^T(x - x^*), \text{ and formula(22)}) \\ &\leq ||x - x^*||_2^2 + t^2||\nabla f(x)||_2^2 - 2t\big(\frac{Lm}{L+m}||x - x^*||_2^2 + \frac{2t}{L+m}||\nabla f(x)||_2^2\big) \\ &= (1 - t\frac{2Lm}{L+m})||x - x^*||_2^2 + t(t - \frac{2}{L+m})||\nabla f(x)||_2^2 \\ &\leq (1 - t\frac{2Lm}{L+m})||x - x^*||_2^2 \end{aligned}$$

Take $x^+ = x^{(k)}, x = x^{(k-1)}$ and $c = 1 - t\frac{2Lm}{L+m}$, it gives:

$$||x^{(k)} - x^*||_2^2 \leq c||x^{(k-1)} - x^*||_2^2 \leq c^2||x^{(k-2)} - x^*||_2^2 \leq c^k||x^{(0)} - x^*||_2^2 \tag{37}$$

Hence, bound on function value (according to formula(13)):

$$f(x^{(k)}) - f^* \leq \frac{L}{2}||x^{(k)} - x^*||_2^2 \leq \frac{c^k L}{2}||x^{(0)} - x^*||_2^2 \tag{38}$$

Therefore, number of iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(log(1/\epsilon))$ □