There exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient. That is, with a good understanding of one, we can easily understand the other one. See `http://xingyuzhou.org/talks/Fenchel_duality.pdf` for a quick summary!

(Better) Convergence Using Strong Convexity

Second Order Conditions for Convexity

Theorem

A twice differential function $f : \mathcal{D} \to \Re$ *for a nonempty open convex set* $\mathcal{D}$

1. *is convex if and only if its domain is convex and its Hessian matrix is positive semidefinite at each point in* $\mathcal{D}$. *That is* $\nabla^2 f(\mathbf{x}) \succeq 0 \quad \forall\, \mathbf{x} \in \mathcal{D}$
2. *is strictly convex if its domain is convex and its Hessian matrix is positive definite at each point in* $\mathcal{D}$. *That is* $\nabla^2 f(\mathbf{x}) \succ 0 \quad \forall\, \mathbf{x} \in \mathcal{D}$
3. *is uniformly convex if and only if its domain is convex and its Hessian matrix is uniformly positive definite at each point in* $\mathcal{D}$. *That is, for any* $\mathbf{v} \in \Re^n$ *and any* $\mathbf{x} \in \mathcal{D}$, *there exists a* $c > 0$ *such that* $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq c||\mathbf{v}||^2$

c and m are used interchangebly as the strong convexity factor/constant

Proof of Second Order Conditions for Convexity

In other words

$$\nabla^2 f(\mathbf{x}) \succeq c I_{n\times n}$$

where $I_{n\times n}$ is the $n \times n$ identity matrix and $\succeq$ corresponds to the positive semidefinite inequality. That is, the function f is strongly convex *iff* $\nabla^2 f(\mathbf{x}) - c I_{n\times n}$ is positive semidefinite, for all $\mathbf{x} \in \mathcal{D}$ and for some constant $c > 0$, which corresponds to the positive minimum curvature of f.

PROOF: We will prove only the first statement; the other two statements are proved in a similar manner.

Necessity: Suppose f is a convex function, and consider a point $\mathbf{x} \in \mathcal{D}$. We will prove that for any $\mathbf{h} \in \Re^n$, $\mathbf{h}^T \nabla^2 f(\mathbf{x}) \mathbf{h} \geq 0$. Since f is convex, we have

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + t\nabla^T f(\mathbf{x})\mathbf{h} \tag{46}$$

Consider the function $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ defined on the domain $\mathcal{D}_\phi = [0, 1]$.

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h}) \frac{dx_i}{dt} = \mathbf{h}^T . \nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2 . \frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h}) \frac{dx_i}{dt} = \mathbf{h}^T . \nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2 . \frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

$$f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{x}) + t\mathbf{h}^T \nabla f(\mathbf{x}) + t^2 \frac{1}{2} h^T \nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3)$$

Proof of Second Order Conditions for Convexity (contd.)

In conjunction with (46), the above equation implies that

$$\frac{t^2}{2} h^T \nabla^2 f(\mathbf{x}) \mathbf{h} + O(t^3) \geq 0$$

Dividing by t^2 and taking limits as $t \to 0$, we get

$$h^T \nabla^2 f(\mathbf{x}) \mathbf{h} \geq 0$$

For necessary condition, take limits

Proof of Second Order Conditions for Convexity (contd.)

Sufficiency: Suppose that the Hessian matrix is positive semidefinite at each point $\mathbf{x} \in \mathcal{D}$. Consider the same function $\phi(t)$ defined above with $\mathbf{h} = \mathbf{y} - \mathbf{x}$ for $\mathbf{y}, \mathbf{x} \in \mathcal{D}$. Applying Taylor's theorem with $n = 2$ and $a = 0$, we obtain,

$$\phi(1) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(c)$$

for some $c \in (0, 1)$. Writing this equation in terms of f gives

$$f(\mathbf{x}) = f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{y})$$

where $\mathbf{z} = \mathbf{y} + c(\mathbf{x} - \mathbf{y})$. Since $\mathcal{D}$ is convex, $\mathbf{z} \in \mathcal{D}$. Thus, $\nabla^2 f(\mathbf{z}) \succeq 0$. It follows that

$$f(\mathbf{x}) \geq f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y})$$

By a previous result, the function f is convex.

Lipschitz Continuity vs. Strong Convexity

- Lipschitz continuity of gradient (references to ∇^2 assume double differentiability)

$$\nabla^2 f(x) \preceq LI$$

$$\left\|\nabla f(x) - \nabla f(y)\right\| \leq L\|x - y\|$$

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

- Strong convexity: Curvature should be **atleast somewhat** positive

$$\nabla^2 f(x) \succeq mI$$

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

  - $m = 0$ corresponds to (sufficient condition for) normal convexity.
  - Later: For example, augmented Lagrangian is used to introduce strong convexity

Conjugate Functions, Strong Convexity and Lipschitz Continuity

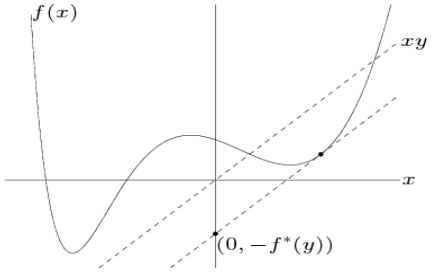
- Conjugate Function of $f : \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x} \in \mathcal{D}} (\mathbf{h}^T \mathbf{x} - f(\mathbf{x}))$
- f^* is convex and closed (even if f is not)
- $\nabla f^*(\mathbf{h}) = \operatorname{argmax}_{\mathbf{x} \in \mathcal{D}} (\mathbf{h}^T \mathbf{x} - f(\mathbf{x}))$
- If f is closed and strongly convex with parameter m, then f^* has a Lipschitz continuous gradient with parameter $1/m$. lower bound on curvature of f is 1/upper bound on curvature of f^*
- If f is convex and has a Lipschitz continuous gradient with parameter L, then f^* is strongly convex with parameter $1/L$

There exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient.

Conjugate function

the **conjugate** of a function f is

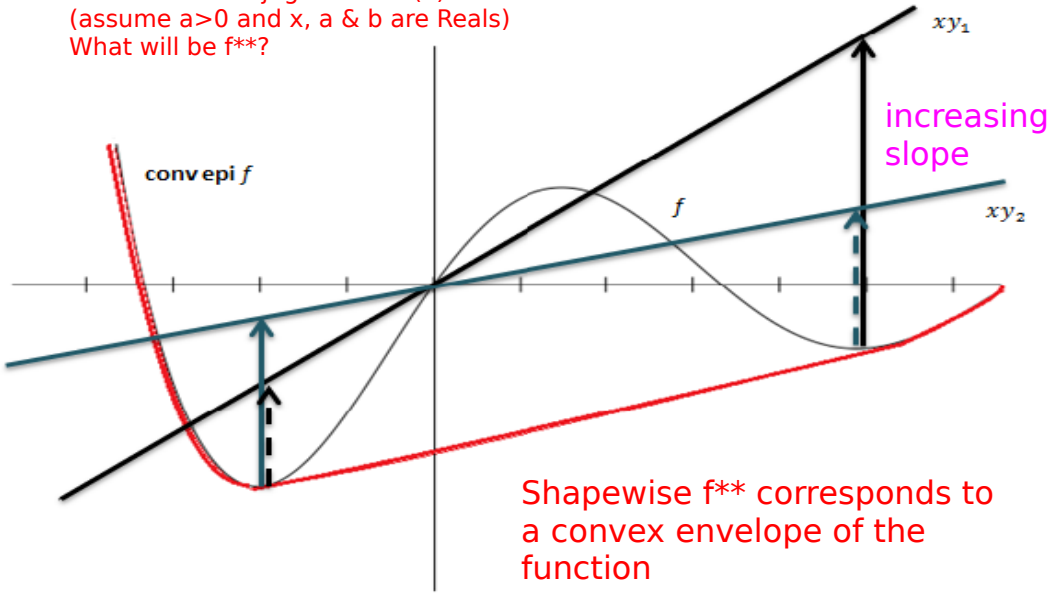
$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$



f^* is convex (even if f is not)

f** is the convex envelope of f

Homework: Find convex conjugate f* of f(x) = a x^2 + bx
(assume a>0 and x, a & b are Reals)
What will be f**?



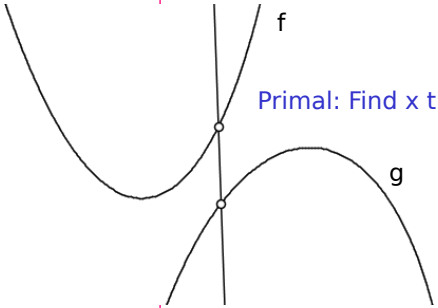
Shapewise f** corresponds to a convex envelope of the function

Fenchel Duality, Strong Convexity and Lipschitz Continuity

- Let f be a closed convex function on $\Re^n$ and let g be a closed concave function on $\Re^n$. Then, under some general conditions:

$$\inf_{\mathbf{x}}(f(\mathbf{x}) - g(\mathbf{x})) = \sup_{\mathbf{h}}(g^*(\mathbf{h}) - f^*(\mathbf{h}))$$

  where f^* is the convex conjugate of f and g^* is the concave conjugate of g
- Thus, there exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient. That is, with a good understanding of one, we can easily understand the other one. See `http://xingyuzhou.org/talks/Fenchel_duality.pdf` for a quick summary!



Primal: Find x that gives smalles gap between f and g

Dual: Find slope h that gives largest gap between g* and f*

Since f and g are convex, the gaps are the same
Otherwise, we expect largest gap in the dual to be less than or equal to gap in primal

Lipschitz Continuity vs. Strong Convexity: Example

- Consider the linear regression loss function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|^2$
- $\nabla f(\mathbf{x}) = -A^T(\mathbf{y} - A\mathbf{x})$
- $\nabla^2 f(\mathbf{x}) = A^T A$
- One can show that

Max and min eigenvalues of A^T A characterize strong convexity and Lipschitz continuity respective.

Lipschitz Continuity vs. Strong Convexity: Example

- Consider the linear regression loss function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|^2$
- $\nabla f(\mathbf{x}) = -A^T(\mathbf{y} - A\mathbf{x})$
- $\nabla^2 f(\mathbf{x}) = A^T A$
- One can show that
  - $\nabla^2 f(\mathbf{x}) = A^T A \preceq LI$ where $L = \sigma_{max}$ is the largest eigenvalue of $A^T A$
  - $\nabla^2 f(\mathbf{x}) = A^T A \succeq mI$ where $m = \sigma_{min}$ is the smallest eigenvalue of $A^T A$

L/m puts some bound on the condition number

Using Strong Convexity: Revisiting Convergence Analysis

- $f(y) \geq f(x) + \nabla^\top f(x)(y-x) + \frac{m}{2}\|y-x\|^2$
  $\geq$ minimum value the RHS can take as a function of y
- Minimum value of RHS
  $\nabla f(x) + my - mx = 0$
  $\implies y = x - \frac{1}{m}\nabla f(x)$
- Thus,
  $f(y) \geq f(x) + \nabla^\top f(x)\left(-\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(x)\right\|^2$
  $\implies f(y) \geq f(x) - \frac{1}{2m}\left\|\nabla f(x)\right\|^2$
  - Here, LHS is independent of x, and RHS is independent of y

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

$$f(x^*) \geq f(x) - \frac{1}{2m}\left\|\nabla f(x)\right\|^2$$

- If $\left\|\nabla f(x)\right\|$ is small, the point is nearly optimal
  - If $\left\|\nabla f(x)\right\| \leq \sqrt{2m\epsilon}$, then:
    $f(x) - f(x^*) \leq \epsilon$
  - As the gradient $\left\|\nabla f(x)\right\|$ approaches 0, we get closer to the optimal solution x^*

- Since f is strongly convex, and also Lipschitz continuous, we have for some L :

$$f(x^{k+1}) \leq f(x^k) + (\frac{Lt^2}{2} - t)\left\|\nabla f(x^k)\right\|^2$$

Recall from previous class

- We also consider
  $0 < t \leq \frac{2}{L}(1 - c_1) \implies \frac{Lt^2}{2} - t \leq -c_1 t$

Most books and notes simplify by assuming c1 = 1/2

- Thus, we get the exit condition of backtracking line search

$$f(x^{k+1}) \leq f(x^k) - c_1 t\left\|\nabla f(x^k)\right\|^2$$

$$\implies f\left(x^k - t\nabla f(x^k)\right) \leq f(x^k) - c_1 t\left\|\nabla f(x^k)\right\|^2$$

- Often $c_1 = \frac{1}{2}$. Convergence of gradient descent, given this condition, has been proved.

- Let $p^* = f(x^*)$
- $f\big(x - t\nabla f(x)\big) \leq f(x) - t\big\|\nabla f(x)\big\|^2 + \frac{Lt^2}{2}\big\|\nabla f(x)\big\|^2$
  - RHS here will be maximum for $t = \frac{1}{L}$

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

  $\implies f\big(x - t^*\nabla f(x)\big) - p^* \leq f(x) - \frac{1}{2L}\big\|\nabla f(x)\big\|^2 - p^*$
- From strong convexity, we had

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

  $\implies p^* \geq f(x) - \frac{1}{2m}\big\|\nabla f(x)\big\|^2$

  $\implies \big\|\nabla f(x)\big\|^2 \geq 2m\left(f(x) - p^*\right)$

Since f(y) was lower bounded by a function of x for all choices of y, we should have that the minimum value of f(y), that is p* is lower bounded by the same function of x

- Thus,

$$f\left(x - t^*\nabla f(x)\right) - p^* \leq f(x) - \frac{1}{2L}\left\|\nabla f(x)\right\|^2 - p^*$$
$$\implies f\left(x - t^*\nabla f(x)\right) - p^* \leq f(x) - \frac{2m}{2L}\left(f(x) - p^*\right) - p^*$$
$$\implies f\left(x - t^*\nabla f(x)\right) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(x) - p^*\right)$$

STEP 1

- Which is,

$$f(x^k) - p^* \leq \left(1 - \frac{m}{L}\right)\left(f(x^{k-1}) - p^*\right)$$
$$\leq \left(1 - \frac{m}{L}\right)^2\left(f(x^{k-2}) - p^*\right)$$
$$\vdots$$
$$\leq \left(1 - \frac{m}{L}\right)^k\left(f(x^{(0)}) - p^*\right)$$

STEP 2

Is this Q-linear convergence... ? ANS: Yes. With r=1-m/L < 1

- We get linear convergence

$$f(x^k) - p^* \leq \left(1 - \frac{m}{L}\right)^k \left(f(x^{(0)}) - p^*\right)$$

  - Here, $\frac{m}{L} \in (0, 1)$
  - This is, loosely speaking, faster than what we got using only Lipschitz continuity, which was: $f(x^k) - p^* \leq \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk}$ (sublinear convergence)

- Called linear convergence because plot of iterations on the x-axis, and difference in the function values on the y-axis on a log scale makes it look linear
- To obtain $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(log(1/\epsilon))$ iterations

Because of the log scale, some people also call this exponential convergence

Summary of Convergence Rate of Gradient Descent Method

For the gradient method, it can be proved that if f is strongly convex,

$$f(\mathbf{x}^{(k)}) - p^* \leq \rho^k \left(f(\mathbf{x}^{(0)} - p^* \right) \tag{47}$$

The value of the linear convergence factor $\rho \in (0, 1)$ depends on the strong convexity constant c, the value of $\mathbf{x}^{(0)}$ and type of ray search employed.

The convergence rate is $1 - m/L$, where L/m is proportional to the *condition number* of the Hessian. Large eigenvalues correspond to high curvature directions and small eigenvalues correspond to low curvature directions. Many methods (such as conjugate gradient) try to improve upon the gradient method by making the hessian better conditioned. Convergence can be very slow even for moderately well-conditioned problems, with condition number in the 100s. is only an $O(n)$ operation. The gradient descent method however works very well if the function is isotropic, that is if the level-curves are spherical or nearly spherical.

The convergence of the steepest descent method can be stated in the same form as in (47), using the fact that any norm can be bounded in terms of the Euclidean norm, *i.e.*, there exists a constant $\eta \in (0, 1]$ such that $||\mathbf{x}|| \geq \eta ||\mathbf{x}||_2$

R-convergence assuming Strong convexity

- Now, let us consider the convergence result we got by assuming Strong convexity with backtracking and exact line searches:

$$f(x^k) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^k \left(f(x^{(0)}) - f(x^*)\right)$$

- Here, v^k can be considered $\left(1 - \frac{m}{M}\right)^k \alpha$
  - $v^* = 0$
- We get

$$\frac{v^{k+1} - v^*}{v^k - v^*} = \left(1 - \frac{m}{M}\right) \in (0, 1)$$

  - We now have an upper bound < 1, unlike before
- As $r = \left(1 - \frac{m}{M}\right) \in (0, 1)$, v^k is Q-linearly convergent
  - Thus, under strong convexity, gradient descent is R-linearly convergent

- *Question:* Is gradient descent under Strong convexity also Q-linearly convergent?
- Recall one of the intermediate steps in getting the convergence results:
  $f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{m}{M}\right)\left(f(x^k) - f(x^*)\right)$
  - $\implies \frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \left(1 - \frac{m}{M}\right)$
- Now, $r = \left(1 - \frac{m}{M}\right) \in (0, 1)$
- Yes, gradient descent under Strong convexity is also Q-linearly convergent

(Sub)Gradient Descent: Generalization of Gradient Descent

Given a convex function $f: \mathbb{R}^n \to \mathbb{R}$, not necessarily differentiable. Subgradient method is just like gradient descent, but replacing gradients with subgradients. *I.e.*, initialize $x^{(0)}$, then repeat

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - t^k \cdot \mathbf{h}^{(k-1)}, k = 1, 2, 3, \cdots$$

where $\mathbf{h}^{(k-1)}$ is **any** subgradient of f at $\mathbf{x}^{(k-1)}$. We keep track of best iterate $\mathbf{x}^k_{best}$ among $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(k)}$:

$$f(\mathbf{x}^{(k)}_{best}) = \min_{i=1,\cdots,k} f(\mathbf{x}^{(i)})$$

Deviation 1: Can't easily ensure necessary condition for descent
So, keep track of best iterate...

To update each $\mathbf{x}^{(i)}$, there are basically two ways to select the step size:

- Fixed step size: $t^k = t$ for all $k = 1, 2, 3 \cdots$
- Diminishing step size: choose t^k to satisfy

Deviation 2: Step sizes not as per Wolfe conditions. Instead individually diminishing, but collectively non-trivial step sizes

$$\lim_{k\to\infty} (t^k) = 0, \quad \sum_{k=1}^{\infty} t^k = \infty$$

1/k, 1/sqrt{k}... give divergent sums
1/k^2 gives a convergent sum (hence unacceptable)

Subgradient Algorithm: Convergence analysis

Given the convex function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies:

- f is Lipschitz continuous with constant $l > 0$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq l||\mathbf{x} - \mathbf{y}|| \text{ for all } \mathbf{x}, \mathbf{y}$$

- $||\mathbf{x}^{(1)} - \mathbf{x}^*|| \leq R$ which means it is bounded Existence of R rather than actual value matters

Theorem

For a fixed step size t, subgradient method satisfies

$$\lim_{k \to \infty} f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + \frac{l^2 t}{2}$$

For diminishing step size such as $t^k = O\left(\frac{1}{\sqrt{k}}\right)$,

$$f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + O\left(\frac{1}{\sqrt{k}}\right)$$

Subgradient Descent: Convergence Analysis (contd.)

Proof:

Try and characterize distance of x^{k+1} and x* in terms of x^k and x*

$$\begin{aligned} ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &= ||\mathbf{x}^{(k)} - t^k \mathbf{h}^{(k)} - \mathbf{x}^*||^2 \\ &= ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k (\mathbf{h}^{(k)})^T (\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2 ||\mathbf{h}^{(k)}||^2 \end{aligned} \quad \text{(a)}$$

By definition of the subgradient method, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T (\mathbf{x}^* - \mathbf{x}^{(k)})$$

$$-(\mathbf{h}^{(k)})^T (\mathbf{x}^* - \mathbf{x}^{(k)}) \leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) \quad \text{(b)}$$

Using this inequality, for $k, k-1, \ldots i, i-1, \ldots 0$ we have

Substituting from (b) in (a) across iterations

Subgradient Descent: Convergence Analysis (contd.)

Proof:

$$\begin{aligned} ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &= ||\mathbf{x}^{(k)} - t^k \mathbf{h}^{(k)} - \mathbf{x}^*||^2 \\ &= ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k (\mathbf{h}^{(k)})^T (\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2 ||\mathbf{h}^{(k)}||^2 \end{aligned}$$

By definition of the subgradient method, we have

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T (\mathbf{x}^* - \mathbf{x}^{(k)}) \\ -(\mathbf{h}^{(k)})^T (\mathbf{x}^* - \mathbf{x}^{(k)}) &\leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) \end{aligned}$$

Using this inequality, for $k, k-1, \ldots i, i-1, \ldots 0$ we have

$$\begin{aligned} ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &\leq ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k (f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) + (t^k)^2 ||\mathbf{h}^{(k)}||^2 \\ &\leq ||\mathbf{x}^{(1)} - \mathbf{x}^*||^2 - 2 \sum_{i=1}^{k} t^i (f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k} (t^i)^2 ||\mathbf{h}^{(i)}||^2 \end{aligned}$$

Subgradient Descent: Convergence Analysis (contd.)

And since this is lower bounded by 0, we have

$$0 \leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 \leq R^2 - 2\sum_{i=1}^{k} t^i (f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k} (t^i)^2 l^2$$

$$\Rightarrow 2\sum_{i=1}^{k} t^i (f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k} (t^i)^2 l^2$$

$$\Rightarrow 2(\sum_{i=1}^{k} t^i)(f(\mathbf{x}^{(k)}_{best}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k} (t^i)^2 l^2$$

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + \rho^2 t^2 k}{2tk} \to \frac{\rho^2 t}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + l^2 t^2 k}{2tk} \to \frac{l^2 t}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

$$\sum_{i=0}^{k} (t^i)^2 \leq 0, \sum_{i=0}^{k} t^i = \infty$$

therefore,

$$\frac{R^2 + l^2 \sum_{i=0}^{k} (t^i)^2}{2 \sum_{i=0}^{k} t^i} \to 0, \text{ as } k \to \infty,$$

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

$$\frac{R^2 + l^2 \sum_{i=0}^{k} (t^i)^2}{2 \sum_{i=0}^{k} t^i} = \frac{Rl}{\sqrt{k}}.$$

That is, subgradient method has convergence rate of $O(\frac{1}{\sqrt{k}})$, and to get $f(x^{(k)}_{best}) - f(x^*) \leq \epsilon$, needs $O(\frac{1}{\epsilon^2})$ iterations.

This is a much worse convergence rate than even $O\left(\frac{1}{k}\right)$ obtained for gradient descent under Lipschitz continuity alone.