More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x, y)$ is convex in (x, y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x, y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})||$, where, $P_C(\mathbf{x}, \mathbf{y}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})}{\|\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})\|}$

More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x, y)$ is convex in (x, y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x, y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})||$, where, $P_C(\mathbf{x}, \mathbf{y}) = \text{argmin} \;\; d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})}{\|\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})\|}$ ....The point of intersection of convex sets C_1, C_2,...C_m by minimizing...

More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x,y)$ is convex in (x,y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x,y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})||$, where, $P_C(\mathbf{x}, \mathbf{y}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})}{\|\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})\|}$ ....The point of intersection of convex sets C_1, C_2,...C_m by minimizing... (Subgradients and Alternating Projections)
  - $\underset{y \in \mathcal{C}}{\text{argmin}}\, d(\mathbf{x}, \mathcal{C})$ is a special case of the proximity operator: $prox_c(\mathbf{x}) = \underset{y}{\text{argmin}}\, PROX_c(\mathbf{x}, \mathbf{y})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}, \mathbf{y}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||$ The special case is when

More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x,y)$ is convex in (x,y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x,y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})||$, where, $P_C(\mathbf{x}, \mathbf{y}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})}{||\mathbf{x} - P_C(\mathbf{x}, \mathbf{y})||}$ ....The point of intersection of convex sets C_1, C_2,...C_m by minimizing... (Subgradients and Alternating Projections)
  - $\underset{y \in \mathcal{C}}{\text{argmin}}\, d(\mathbf{x}, \mathcal{C})$ is a special case of the proximity operator: $prox_c(\mathbf{x}) = \underset{y}{\text{argmin}}\, PROX_c(\mathbf{x}, \mathbf{y})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}, \mathbf{y}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||$ The special case is when $c(\mathbf{y})$ is the indicator function $I_C(\mathbf{y})$ introduced earlier to eliminate the contraints of an optimization problem.
    - ★ Recall that $\partial I_C(\mathbf{y}) = N_C(\mathbf{y}) = \{\mathbf{h} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
    - ★ The subdifferential $\partial PROX_c(\mathbf{x}, \mathbf{y}) = \partial c(\mathbf{y}) + \mathbf{y} - \mathbf{x}$ which can now be obtained for the special case $c(\mathbf{y}) = I_C(\mathbf{y})$.

More Subgradient Calculus: Perspective (Advanced)

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Perspective Function:** The perspective of a function $f: \Re^n \to \Re$ is the function $g: R^n \times \Re \to \Re$, $g(x,t) = tf(x/t)$. Function g is convex if f is convex on $\mathbf{dom} g = \left\{(x,t)|x/t \in \mathbf{dom} f, t > 0\right\}$. For example,
  - The perspective of $f(x) = x^T x$ is (quadratic-over-linear) function $g(x,t) = \frac{x^T x}{t}$ and is convex.
  - The perspective of negative logarithm $f(x) = -\log x$ is the relative entropy function $g(x,t) = t\log t - t\log x$ and is convex.

relative to t

Illustrating the Why and How of (Sub)Gradient on Lasso

Recap: Subgradients for the 'Lasso' Problem in Machine Learning

Recall Lasso ($\min_{\mathbf{x}} f(\mathbf{x})$) as an example to illustrate subgradients of affine composition:

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

y is fixed

The subgradients of $f(\mathbf{x})$ are

x - y + lambda s

s.t: s_i = sign(x_i) if x_i != 0
o/w: 0 <= s_i <= 1

Recap: Subgradients for the 'Lasso' Problem in Machine Learning

Recall Lasso ($\min_{\mathbf{x}} f(\mathbf{x})$) as an example to illustrate subgradients of affine composition:

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

The subgradients of $f(\mathbf{x})$ are

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.

results from convex hull of union of subdifferentia

Here we only see "HOW" to compute the subdifferential.

Subgradients in a Lasso sub-problem: Sufficient Condition Test

Invoking "WHY" of subdiff.

We illustrate the sufficient condition again using a sub-problem in Lasso as an example.
Consider the simplified Lasso problem (which is a sub-problem in Lasso):

min_x

$$\cancel{f(\mathbf{x})} = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

Recall the subgradients of $f(\mathbf{x})$:

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.
A solution to this problem is

x_i = 0 if y_i is between -\lambda and \lambda
and there exists an s_i between -1 and +1 for this case

In fact this s_i = y_i / lambda

Subgradients in a Lasso sub-problem: Sufficient Condition Test

We illustrate the sufficient condition again using a sub-problem in Lasso as an example. Consider the simplified Lasso problem (which is a sub-problem in Lasso):

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

Recall the subgradients of $f(\mathbf{x})$:

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.

A solution to this problem is $\mathbf{x}^* = S_\lambda(\mathbf{y})$, where $S_\lambda(\mathbf{y})$ is the **soft-thresholding operator**:

$$S_\lambda(\mathbf{y}) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

Now if $\mathbf{x}^* = S_\lambda(\mathbf{y})$ then there exists a $\mathbf{h}(\mathbf{x}) = 0$. Why? If $y_i > \lambda$, we have $x_i^* - y_i = -\lambda + \lambda \cdot 1 = 0$. The case of $y_i < \lambda$ is similar. If $-\lambda \leq y_i \leq \lambda$, we have $x_i^* - y_i = -y_i + \lambda(\frac{y_i}{\lambda}) = 0$. Here, $s_i = \frac{y_i}{\lambda}$.

Proximal Operator and Sufficient Condition Test

- Recap: $d(\mathbf{x}, \mathcal{C})$ returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}||^2$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function.
- Recap: $\underset{y \in \mathcal{C}}{\text{argmin}}\, ||\mathbf{x} - \mathbf{y}||$ is a special case of the proximal operator: $prox_c(\mathbf{x}) = \underset{y}{\text{argmin}}\, PROX_c(\mathbf{x}, \mathbf{y})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}, \mathbf{y}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||^2$ The special case is when $c(\mathbf{y})$ is the indicator function $I_C(\mathbf{y})$ introduced earlier to eliminate the contraints of an optimization problem.
  - Recall that $\partial I_C(\mathbf{y}) = N_C(\mathbf{y}) = \{\mathbf{h} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
  - For the special case $c(\mathbf{y}) = I_C(\mathbf{y})$, the subdifferential $\partial PROX_c(\mathbf{x}, \mathbf{y}) = \partial c(\mathbf{y}) + \mathbf{y} - \mathbf{x} = \{\mathbf{h} - \mathbf{x} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
  - As per sufficient condition for minimum for this special case, $prox_c(\mathbf{x}) =$ that y in C that is closest to x

Proximal Operator and Sufficient Condition Test

- Recap: $d(\mathbf{x}, \mathcal{C})$ returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}||^2$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function.
- Recap: $\underset{y \in \mathcal{C}}{\text{argmin}} ||\mathbf{x} - \mathbf{y}||$ is a special case of the proximal operator: $prox_c(\mathbf{x}) = \underset{y}{\text{argmin}}\, PROX_c(\mathbf{x}, \mathbf{y})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}, \mathbf{y}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||^2$ The special case is when $c(\mathbf{y})$ is the indicator function $I_C(\mathbf{y})$ introduced earlier to eliminate the contraints of an optimization problem.
    - Recall that $\partial I_C(\mathbf{y}) = N_C(\mathbf{y}) = \{\mathbf{h} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
    - For the special case $c(\mathbf{y}) = I_C(\mathbf{y})$, the subdifferential $\partial PROX_c(\mathbf{x}, \mathbf{y}) = \partial c(\mathbf{y}) + \mathbf{y} - \mathbf{x} = \{\mathbf{h} - \mathbf{x} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
    - As per sufficient condition for minimum for this special case, $prox_c(\mathbf{x}) = \underset{y \in \mathcal{C}}{\text{argmin}} ||\mathbf{x} - \mathbf{y}||$

that y in C that is closest to x

- We will invoke this when we discuss the **proximal gradient descent** algorithm

Convexity by Restriction to line, (Sub)Gradients and Monotonicity

Convexity by Restricting to Line

A useful technique for verifying the convexity of a function is to investigate its convexity, by restricting the function to a line and checking for the convexity of a function of single variable.

Theorem

A function $f : \mathcal{D} \rightarrow \Re$ is (strictly) convex if and only if the function $\phi : \mathcal{D}_\phi \rightarrow \Re$ defined below, is (strictly) convex in t for every $\mathbf{x} \in \Re^n$ and for every $\mathbf{h} \in \Re^n$ Direction vector or line

Here we see connection with direction, independent of differentiability

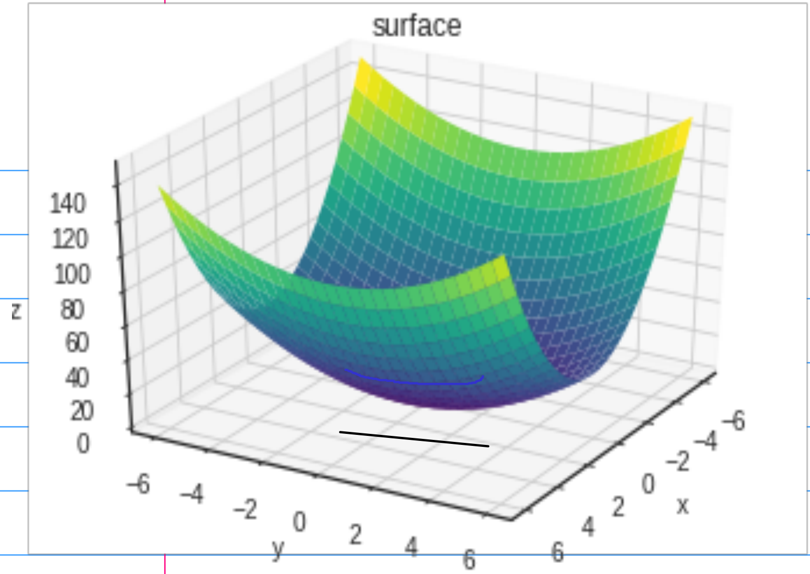
$$\phi(t) = f(\mathbf{x} + t\mathbf{h})$$

with the domain of ϕ given by $\mathcal{D}_\phi = \{t | \mathbf{x} + t\mathbf{h} \in \mathcal{D}\}$.

We saw the connection with R: convex differentiable fn iff directional deriv is convex along every direction

Thus, we have see that

- If a function has a local optimum at $\mathbf{x}^*$, it as a local optimum along each component x_i^* of $\mathbf{x}^*$
- If a function is convex in $\mathbf{x}$, it will be convex in each component x_i of $\mathbf{x}$


surface
z
140
120
100
80
60
40
20
0
-6
-4
-2
0
2
4
6
y
x

Convexity by Restricting to Line (contd.)

Proof: We will prove the necessity and sufficiency of the convexity of ϕ for a convex function f. The proof for necessity and sufficiency of the strict convexity of ϕ for a strictly convex f is very similar and is left as an exercise.

Proof of Necessity: Assume that f is convex. And we need to prove that $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is also convex. Let $t_1, t_2 \in \mathcal{D}_\phi$ and $\theta \in [0, 1]$. Then, (for any direction h)

$$\phi(\theta t_1 + (1-\theta)t_2) = f\Big(\theta(\mathbf{x} + t_1\mathbf{h}) + (1-\theta)(\mathbf{x} + t_2\mathbf{h})\Big)$$

<= \theta f(...x_1) + (1-\theta) f(...x_2)

Convexity by Restricting to Line (contd.)

Proof: We will prove the necessity and sufficiency of the convexity of ϕ for a convex function f. The proof for necessity and sufficiency of the strict convexity of ϕ for a strictly convex f is very similar and is left as an exercise.

Proof of Necessity: Assume that f is convex. And we need to prove that $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is also convex. Let $t_1, t_2 \in \mathcal{D}_\phi$ and $\theta \in [0, 1]$. Then,

$$\begin{aligned} \phi(\theta t_1 + (1-\theta)t_2) &= f\big(\theta(\mathbf{x} + t_1\mathbf{h}) + (1-\theta)(\mathbf{x} + t_2\mathbf{h})\big) \\ &\leq \theta f\big((\mathbf{x} + t_1\mathbf{h})\big) + (1-\theta) f\big((\mathbf{x} + t_2\mathbf{h})\big) = \theta\phi(t_1) + (1-\theta)\phi(t_2) \end{aligned} \tag{16}$$

Thus, ϕ is convex.

Convexity by Restricting to Line (contd.)

Proof of Sufficiency: Assume that for every $\mathbf{h} \in \Re^n$ and every $\mathbf{x} \in \Re^n$, $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is convex. We will prove that f is convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$. Take, $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{h} = \mathbf{x}_2 - \mathbf{x}_1$. We know that $\phi(t) = f\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)$ is convex, with $\phi(1) = f(\mathbf{x}_2)$ and $\phi(0) = f(\mathbf{x}_1)$. Therefore, for any $\theta \in [0, 1]$

$$f\big(\theta \mathbf{x}_2 + (1-\theta)\mathbf{x}_1\big) = \phi(\theta)$$

<= theta \phi(1) + (1-theta)\phi(0)
= theta f(x2) + (1-theta) f(x1)

Convexity by Restricting to Line (contd.)

Proof of Sufficiency: Assume that for every $\mathbf{h} \in \Re^n$ and every $\mathbf{x} \in \Re^n$, $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is convex. We will prove that f is convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$. Take, $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{h} = \mathbf{x}_2 - \mathbf{x}_1$. We know that $\phi(t) = f\left(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\right)$ is convex, with $\phi(1) = f(\mathbf{x}_2)$ and $\phi(0) = f(\mathbf{x}_1)$. Therefore, for any $\theta \in [0, 1]$

$$\begin{aligned} \underline{f\left(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1\right)} &= \phi(\theta) \\ \leq \theta\phi(1) + (1-\theta)\phi(0) \leq \underline{\theta f(\mathbf{x}_2) + (1-\theta)f(\mathbf{x}_1)} & \end{aligned} \tag{17}$$

This implies that f is convex.

More on SubGradient kind of functions: Monotonicity

A differentiable function $f : \Re \rightarrow \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f : \Re^n \rightarrow \Re$?

Ans: Yes. We need a notion of monotonicity of vectors (subgradients)

More on SubGradient kind of functions: Monotonicity

A differentiable function $f: \Re \to \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f: \Re^n \to \Re$? View subgradient as an instance of a general function $\mathbf{h}: \mathcal{D} \to \Re^n$ and $\mathcal{D} \subseteq \Re^n$. Then

h is monotone iff the dot product of h(x) - h(y) with x-y is non-negative for all x and y

The component-wise notion of monotonicity of a vector h is a special case of the above more general monotonicity

More on SubGradient kind of functions: Monotonicity

A differentiable function $f : \Re \to \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f : \Re^n \to \Re$? View subgradient as an instance of a general function $\mathbf{h} : \mathcal{D} \to \Re^n$ and $\mathcal{D} \subseteq \Re^n$. Then

Definition

1. $\mathbf{h}$ is *monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$,

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) \geq 0 \tag{18}$$

The component-wise notion of monotonicity of a vector h is a special case of the above more general monotonicity

More on SubGradient kind of functions: Monotonicity (contd)

Definition

2. $\mathbf{h}$ is *strictly monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$ with $\mathbf{x}_1 \neq \mathbf{x}_2$,

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) > 0 \tag{19}$$

3. $\mathbf{h}$ is *uniformly* or *strongly monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, there is a constant $c > 0$ such that

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) \geq c||\mathbf{x}_1 - \mathbf{x}_2||^2 \tag{20}$$

Several other notions of uniform monotonicity can be motivated by simply looking at other lower bounds (instead of this quadratic L2 norm based lower bound)

Several such lower bounds are some divergence functions between x1 and x2

(Sub)Gradients and Convexity

Based on the definition of monotonic functions, we next show the relationship between convexity of a function and monotonicity of its (sub)gradient:

Theorem

Let $f : \mathcal{D} \to \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,

1. *f is convex on $\mathcal{D}$* **iff** *its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$:* $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq 0$
2. *f is strictly convex on $\mathcal{D}$* **iff** *its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$:* $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) > 0$
3. *f is uniformly or strongly convex on $\mathcal{D}$* **iff** *its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$,* $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2$ *for some constant $c > 0$.*

While these results also hold for subgradients $\mathbf{h}$, we will show them only for gradients ∇f
For proving the equivalence, we invoke the \phi defined previously as well as mean value theorem etc on \phi

(Sub)Gradients and Convexity (contd)

Proof:

Necessity: Suppose f is strongly convex on $\mathcal{D}$. Then we know from an earlier result that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \underline{\nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})} - \frac{1}{2}c||\mathbf{y}-\mathbf{x}||^2$$

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \underline{\nabla^T f(\mathbf{y})(\mathbf{x}-\mathbf{y})} - \frac{1}{2}c||\mathbf{x}-\mathbf{y}||^2$$

Adding the two inequalities,

(Sub)Gradients and Convexity (contd)

Proof:
Necessity: Suppose f is strongly convex on $\mathcal{D}$. Then we know from an earlier result that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) - \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x} - \mathbf{y}) - \frac{1}{2}c||\mathbf{x} - \mathbf{y}||^2$$

Adding the two inequalities, we get uniform/strong monotonicity in definition (3). If f is convex, the inequalities hold with $c = 0$, yielding monotonicity in definition (1). If f is strictly convex, the inequalities will be strict, yielding strict monotonicity in definition (2).

(Sub)Gradients and Convexity (contd)

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big)$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

(Sub)Gradients and Convexity (contd)

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big)$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

$$\phi(1) - \phi(0) = \phi'(t) \tag{21}$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$, (21) translates to

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \tag{22}$$

Also, by definition of monotonicity of ∇f,

$$\big(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\big)^T (\mathbf{y} - \mathbf{x}) = \frac{1}{t}\big(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\big)^T (\mathbf{z} - \mathbf{x}) \geq 0 \tag{23}$$

(Sub)Gradients and Convexity (contd)

Combining (22) with (23), we get,

$$f(\mathbf{y}) - f(\mathbf{x}) = \big(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\big)^T (\mathbf{y} - \mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \\ \geq \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \tag{24}$$

By a previous foundational result, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (23) inherited from strict monotonicity, and letting the strict inequality follow through to (24).

(Sub)Gradients and Convexity (contd)

For the case of strong convexity, we have Some more additional work for strong convexity

$$\phi'(t) - \phi'(0) = \left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{y} - \mathbf{x})$$

$$= \frac{1}{t}\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{z} - \mathbf{x}) \geq \frac{1}{t} c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \quad (25)$$

Therefore,

(Sub)Gradients and Convexity (contd)

For the case of strong convexity, we have

$$\phi'(t) - \phi'(0) = \left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{y} - \mathbf{x})$$
$$= \frac{1}{t}\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{z} - \mathbf{x}) \geq \frac{1}{t} c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \quad (25)$$

Therefore,

integrating over this inequality from
t = 0 to t = 1

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)]dt \geq \underline{\frac{1}{2} c||\mathbf{y} - \mathbf{x}||^2} \quad (26)$$

which translates to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2} c||\mathbf{y} - \mathbf{x}||^2$$

Thus, f must be strongly convex.

Descent Algorithms

Some insights on why descent algorithms (based on subgradients for example) will behave well on convex functions

1) Vanishing of subgradient is a sufficient condition for minimization of a convex fn

==> This is handy for constrained optimization as well

2) If f is convex and differentiable, the subgradient is unique = gradient.. In general the convergence rates using gradient are much better than those using subgradients

3) For a convex fn, any point of local min is a point of global min

4) (Sub)gradients exhibit some monotonic behaviour when the function is convex

Descent Algorithms for Optimizing Unconstrained Problems

Techniques relevant for most (convex) optimization problems that do not yield themselves to closed form solutions. We will start with unconstrained minimization.

$$\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$$

For analysis:

- Assume that f is convex and differentiable and that it attains a finite optimal value p^*.
- Minimization techniques produce a sequence of points $\mathbf{x}^{(k)} \in \mathcal{D}$, $k = 0, 1, \ldots$ such that $f\left(\mathbf{x}^{(k)}\right) \to p^*$ as $k \to \infty$ or, $\nabla f\left(\mathbf{x}^{(k)}\right) \to \mathbf{0}$ as $k \to \infty$.
- Iterative techniques for optimization, further require a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$ and sometimes that *epi*(f) is closed. The *epi*(f) can be inferred to be closed either if $\mathcal{D} = \Re^n$ or $f(\mathbf{x}) \to \infty$ as $\mathbf{x} \to \partial\mathcal{D}$. The function $f(x) = \frac{1}{x}$ for $x > 0$ is an example of a function whose *epi*(f) is not closed.