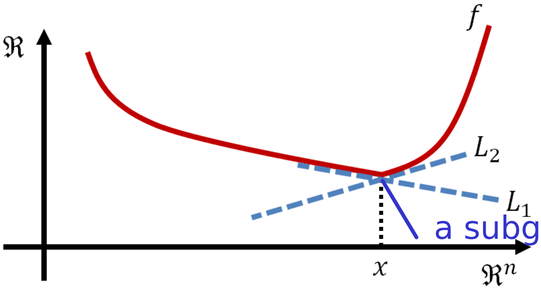
Convexity, Local and Global Optimality, etc.

(Sub)Gradients and Convexity (contd)



a subgradient is a function of this point

In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Recall that a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{y}$$

Thus, intuitively, if a function is differentiable at a point $\mathbf{x}$ then it has a unique subgradient at that point ($\bigtriangledown f(\mathbf{x})$). Formal Proof?

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- Often an indicator function, $I_C : \Re^n \mapsto \Re$, is employed to remove the contraints of an optimization problem (note that convex set $C \subseteq \Re^n$):

$$\min_{\mathbf{x} \in C} f(\mathbf{x}) \iff \min_{\mathbf{x}} f(\mathbf{x}) + I_C(\mathbf{x}), \quad \text{where} \quad I_C(\mathbf{x}) = I\{\mathbf{x} \in C\} = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ \infty & \text{if } \mathbf{x} \notin C \end{cases}$$

  The subdifferential of the indicator function at x is cone spanned by vectors orthogonal to the tangential hyperplanes at x

  (for x on boundary of C)

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- Often an indicator function, $I_C : \Re^n \mapsto \Re$, is employed to remove the contraints of an optimization problem (note that convex set $C \subseteq \Re^n$):

$$\min_{\mathbf{x} \in C} f(\mathbf{x}) \iff \min_{\mathbf{x}} f(\mathbf{x}) + I_C(\mathbf{x}), \quad \text{where} \quad I_C(\mathbf{x}) = I\{\mathbf{x} \in C\} = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ \infty & \text{if } \mathbf{x} \notin C \end{cases}$$

  The subdifferential of the indicator function at x is known as the **normal cone**, $N_C(\mathbf{x})$, of C:

$$N_C(\mathbf{x}) = \partial I_C(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h}^T x \geq \mathbf{h}^T \mathbf{y} \text{ for any } \mathbf{y} \in C\}$$

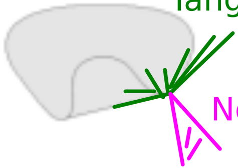
Cone formed by all normals to tangent hyperplanes

Normal Cones (Tangent Cone and Polar) for some Convex Sets

If C is a convex set and if..

- $\mathbf{x} \in int(C)$ then $N_C(\mathbf{x}) = \{\mathbf{0}\}$. In general, if $\mathbf{x} \in int(domain(f))$ then $\partial f(\mathbf{x})$ is nonempty and bounded. By definition, it is an intersection of half spaces
- $\mathbf{x} \in C$ then $N_C(\mathbf{x})$ is a closed convex cone. In general, $\partial f(\mathbf{x})$ is (possibly empty) closed convex set since it is the intersection of half spaces
- There is a relation between the intuitive **tangent cone** and **normal cone** at a point $\mathbf{x} \in \partial C$....This relation is the polar relation.

Let us construct the **normal cone**, $N_C(\mathbf{x})$ for some points in a convex set C:



Normal cone = Polar of tangent cone

Intuition only:Polar corresponds to finding orthogonal vector to tangent plane

First peek into subgradient calculus: Function Convexity First

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

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\}$ is max of LHs <= max of RHS <= sum of max of individual components of RHS

First peek into subgradient calculus: Function Convexity First

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

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\}$ is also convex. [Recall Quiz 1, Question 2]. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is

f(x) = max of dot product of a permutation vector of r 1's with x

First peek into subgradient calculus: Function Convexity First

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

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\}$ is also convex. [Recall Quiz 1, Question 2]. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup\limits_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is

  Uncountably/countably infinite extension of the previous result

  In general, induced matrix norms invoke supremum over vector norms and are therefore convex

First peek into subgradient calculus: Function Convexity First

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

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\}$ is also convex. [Recall Quiz 1, Question 2]. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup\limits_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is convex. For example:
  - The function that returns the maximum eigenvalue of a symmetric matrix X, *viz.*, $\lambda_{max}(X) = \sup\limits_{\mathbf{y} \in \mathcal{S}} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is therefore convex

First peek into subgradient calculus: Function Convexity First

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

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\right\}$ is also convex. [Recall Quiz 1, Question 2]. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is convex. For example:
  - The function that returns the maximum eigenvalue of a symmetric matrix X, *viz.*, $\lambda_{max}(X) = \sup_{\mathbf{y} \in \mathcal{S}} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is a convex function of the symmetrix matrix X.

Basic Subgradient Calculus: Illustration for pointwise Maximum

- Finite pointwise maximum: if $f(\mathbf{x}) = max_{i=1...m} f_i(\mathbf{x})$, then

$\partial f(\mathbf{x}) =$ Find all i such that fi(x) = f(x)
Compute the subdifferential of fi(x)
Take union of these subdifferentials
Find convex hull of the union

Basic Subgradient Calculus: Illustration for pointwise Maximum

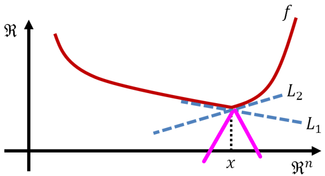
- Finite pointwise maximum: if $f(\mathbf{x}) = max_{i=1...m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup_{i: f_i(\mathbf{x}) = f(\mathbf{x})} \partial f_i(\mathbf{x}) \right)$, which is the convex hull of union of subdifferentials of all active functions at x.
- General pointwise maximum: if $f(\mathbf{x}) = max_{s \in S} f_s(\mathbf{x})$, then under some regularity conditions (on S, f_s), $\partial f(\mathbf{x}) =$

(could also generalize to supremum?)

Find all s such that fs(x) = f(x)
Compute the subdifferential of fs(x)
Take union of these subdifferentials
Find convex hull of the union
Find the closure of this hull

Basic Subgradient Calculus: Illustration for pointwise Maximum

- Finite pointwise maximum: if $f(\mathbf{x}) = max_{i=1\ldots m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup_{i: f_i(\mathbf{x}) = f(\mathbf{x})} \partial f_i(\mathbf{x}) \right)$, which is the convex hull of union of subdifferentials of all active functions at x.
- General pointwise maximum: if $f(\mathbf{x}) = max_{s \in S} f_s(\mathbf{x})$, then under some regularity conditions (on S, f_s), $\partial f(\mathbf{x}) = cl\left\{ conv\left(\bigcup_{s: f_s(\mathbf{x}) = f(\mathbf{x})} \partial f_s(\mathbf{x}) \right) \right\}$



Subgradient of $\|\mathbf{x}\|_1$

Assume $\mathbf{x} \in \Re^n$. Then

- $\|\mathbf{x}\|_1 =$ |x1| + |x2|....+ |xn| = max over 2^n permutation vectors s of x^T s

Subgradient of $\|\mathbf{x}\|_1$

Assume $\mathbf{x} \in \Re^n$. Then

- $\|\mathbf{x}\|_1 = \max_{\mathbf{s} \in \{-1,+1\}^n} \mathbf{x}^T\mathbf{s}$ which is a pointwise maximum of 2^n functions
- Let $\mathcal{S}^* \subseteq \{-1,+1\}^n$ be the set of $\mathbf{s}$ such that for each $\mathbf{s} \in \mathcal{S}^*$, the value of $\mathbf{x}^T\mathbf{s}$ is the same max value.
- Thus, $\partial\|\mathbf{x}\|_1 = conv\left(\bigcup_{\mathbf{s} \in \mathcal{S}^*} \mathbf{s}\right)$.

S* will contain more than a single s only if vector x has some 0's in it
For |x|, at x=0, the subdifferential is [-1,+1]: the closed interval is a convex hull

Differentiable convex function has unique subgradient: Proof

Stated inquitively earlier. Now formally:

Let $f: \Re^n \to \Re$ be a convex function. If f is differentiable at $\mathbf{x} \in \Re^n$ then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$

- We know from (8) that for a differentiable $f: \mathcal{D} \to \Re$ and open convex set $\mathcal{D}$, f is convex **iff**,

  f(y) >= f(x) + < \grad f(x) , (y-x) >

  Thus: \grad f(x) is an element of the subdifferential of f(x)

Differentiable convex function has unique subgradient: Proof

Stated inquitively earlier. Now formally:

Let $f: \Re^n \to \Re$ be a convex function. If f is differentiable at $\mathbf{x} \in \Re^n$ then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$

- We know from (8) that for a differentiable $f: \mathcal{D} \to \Re$ and open convex set $\mathcal{D}$, f is convex **iff**, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$
  Thus, $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$.
- Let $\mathbf{h} \in \partial f(\mathbf{x})$, then $\mathbf{h}^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x})$. Since f is differentiable at $\mathbf{x}$, we have that

We can compute the directional derivative of f along y-x

Differentiable convex function has unique subgradient: Proof

Stated inquitively earlier. Now formally:

Let $f: \Re^n \to \Re$ be a convex function. If f is differentiable at $\mathbf{x} \in \Re^n$ then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$

- We know from (8) that for a differentiable $f: \mathcal{D} \to \Re$ and open convex set $\mathcal{D}$, f is convex **iff**, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$
  Thus, $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$.
- Let $\mathbf{h} \in \partial f(\mathbf{x})$, then $\mathbf{h}^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x})$. Since f is differentiable at $\mathbf{x}$, we have that $\lim_{\mathbf{y} \to \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} = 0$ Letting ||y-x|| shrink
- Thus for any $\epsilon > 0$ there exists a $\delta > 0$ such that $\left| \frac{f(\mathbf{y}) - f(\mathbf{x}) - \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \right| < \epsilon$ whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$.
- Multiplying both sides by $\|\mathbf{y} - \mathbf{x}\|$ and adding $\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$ to both sides, we get $f(x) - f(\bar{x}) < \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \epsilon \|\mathbf{y} - \mathbf{x}\|$ whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$

Differentiable convex function has unique subgradient: Proof

- But then, given that $\mathbf{h} \in \partial f(\mathbf{x})$,

Differentiable convex function has unique subgradient: Proof

- But then, given that $\mathbf{h} \in \partial f(\mathbf{x})$,we obtain
  $\mathbf{h}^T(\mathbf{y}-\mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}) < \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Rearranging we get $(\mathbf{h} - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) < \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Consider $\mathbf{y}-\mathbf{x} =$ some value such that we can eliminate delta eventually (after substituting in the inequality)

Differentiable convex function has unique subgradient: Proof

- But then, given that $\mathbf{h} \in \partial f(\mathbf{x})$,we obtain $\mathbf{h}^T(\mathbf{y}-\mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}) < \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Rearranging we get $(\mathbf{h} - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) < \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Consider $\mathbf{y}-\mathbf{x} = \frac{\delta(\mathbf{h}-\nabla f(\mathbf{x}))}{2\|\mathbf{h}-\nabla f(\mathbf{x})\|}$ that has norm $\|.\| = \frac{\delta}{2}$ less than δ. Then, substituting in the previous step: $(\mathbf{h} - \nabla f(\mathbf{x}))^T\left(\frac{\delta(\mathbf{h}-\nabla f(\mathbf{x}))}{2\|\mathbf{h}-\nabla f(\mathbf{x})\|}\right) < \epsilon\frac{\delta}{2}$
- Canceling out common terms and evaluating dot product as eucledian norm we get: $\|\mathbf{h} - \nabla f(\mathbf{x}))\| < \epsilon$, which should be true for any $\epsilon > 0$, it should be that $\|\mathbf{h} - \nabla f(\mathbf{x}))\| = 0$. Thus, it must be that $\mathbf{h} = \nabla f(\mathbf{x}))$

More Subgradient Calculus: Function Convexity first

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

- **Nonnegative weighted sum:** $f = \sum_{i=1}^{n} \alpha_i f_i$ is convex if each f_i for $1 \leq i \leq n$ is convex and $\alpha_i \geq 0, 1 \leq i \leq n$.
- **Composition with affine function:** $f(Ax + b)$ is convex if f is convex. For example:
  - The log barrier for linear inequalities, $f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$, is convex since $-\log(x)$ is convex.
  - Any norm of an affine function, $f(x) = ||Ax + b||$, is convex.

since 1/(x^2) is > 0

More of Basic Subgradient Calculus

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$
- Affine composition: if $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = A^T \partial f(A\mathbf{x} + b)$
- Norms: important special case, $f(\mathbf{x}) = ||\mathbf{x}||_p$

More of Basic Subgradient Calculus

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$
- Affine composition: if $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = A^T \partial f(A\mathbf{x} + b)$
- Norms: important special case, $f(\mathbf{x}) = ||\mathbf{x}||_p \max_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then

More of Basic Subgradient Calculus

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$
- Affine composition: if $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = A^T \partial f(A\mathbf{x} + b)$
- Norms: important special case, $f(\mathbf{x}) = ||\mathbf{x}||_p \max\limits_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then
$$\partial f(\mathbf{x}) = \left\{\mathbf{y} : ||\mathbf{y}||_q \leq 1 \text{ and } \mathbf{y}^T x = \max_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x}\right\}$$

Subgradients for Lasso

We use Lasso as an example to illustrate subgradients of affine composition:

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

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

Subgradients for Lasso

We use Lasso as an example to illustrate subgradients of affine composition:

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

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 $x_i = 0$.