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 $1 \le i \le n$ is convex and $\alpha_i \ge 0, 1 \le i \le n$.
- **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.

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, $f(\mathbf{x}) = ||\mathbf{x}||_p$

The derivations done in class could be used to show that if any other subgradient exists for g outside the stated set above, that could be used to construct a subgradient for f outside the stated set above as well!

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

On the board we have used y instead of z

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, $f(\mathbf{x}) = ||\mathbf{x}||_p = \max\limits_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then $\partial f(\mathbf{x}) = \left\{ \mathbf{y} : ||\mathbf{y}||_q \le 1 \text{ and } \mathbf{y}^T x = \max\limits_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x} \right\} =$

y corresponds to z where the max is attained

The part above is largely connected to previous discussion on max of convex functions

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, $f(\mathbf{x}) = ||\mathbf{x}||_p = \max\limits_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then

This is derived in class

$$\partial f(\mathbf{x}) = \left\{ \mathbf{y} : ||\mathbf{y}||_q \le 1 \text{ and } \mathbf{y}^T x = \max_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x} \right\} = \underline{\left\{ \mathbf{y} : ||\mathbf{y}||_q \le 1 \text{ and } \mathbf{y}^T x = ||\mathbf{x}||_p \right\}}$$

Why ||y||_q <= 1
is because of Minkowski's
inequality

Subgradients for the 'Lasso' Problem in Machine Learning

We use 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

x - y + \lambda s
Where s = {+1,-1}^n
such that ||x||_1 = s^T x

Subgradients for the 'Lasso' Problem in Machine Learning

We use 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 $x_i = 0$.

Second component is a result of the convex hull

More Subgradient Calculus: Composition

Following functions, though convex, may not be differentiable everywhere. How does one compute their subgradients? (what holds for subgradient also holds for gradient)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument

We will consider only the first case

More Subgradient Calculus: Composition

Following functions, though convex, may not be differentiable everywhere. How does one compute their subgradients? (what holds for subgradient also holds for gradient)

- **Composition with functions:** Let $p : \Re^k \to \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \to \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument

  In both conditions, composition will be concave if p is concave

  Some examples illustrating this property are:
  - $exp\ q(\mathbf{x})$ is convex if q is convex

    exp is a monotonic and convex p
  - $\sum_{i=1}^{m} \log q_i(\mathbf{x})$ is concave if q_i are concave and positive

    p is concave and hence the composition is concave
  - $\log \sum_{i=1}^{m} \exp q_i(\mathbf{x})$ is convex if q_i are convex
  - $1/q(\mathbf{x})$ is convex if q is concave and positive

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \to \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{d}om\ h$ and $q : \Re^n \to \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = \underline{p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y} - \mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y} - \mathbf{x})\right)}$
    Where $\underline{\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})}$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.

p applied to qi(x) is >= p applied to the lower bounds on qi(x)

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.
  - $p\left(q_1(\mathbf{x}) + \underline{\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x})}, \ldots, q_k(\mathbf{x}) + \underline{\mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})}\right) \geq$
    $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + \mathbf{h}_p^T\left(\underline{\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})}\right)$
    Where $\underline{\mathbf{h}_p \in \partial p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right)}$

All we need to do next is club together h_p and h_q and leave only (y-x) in the second component

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.
  - $p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right) \geq$
    $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + \mathbf{h}_p^T\left(\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_p \in \partial p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right)$
  - $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + h_p^T\left(h_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, h_{q_k}^T(\mathbf{y}-\mathbf{x})\right) = f(\mathbf{x}) + \underline{\sum_{i=1}^{k}(h_p)_i h_{q_i}^T}(\mathbf{y}-\mathbf{x})$

That is, $\sum_{i=1}^{k}(h_p)_i h_{q_i}$ is a subgradient of the composite function at $\mathbf{x}$.

$\boxed{\text{H/W: Derive the subdifferentials to example functions on previous slide}}$

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 $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x})||$, where, $P_C(\mathbf{x}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x})}{\|\mathbf{x} - P_C(\mathbf{x})\|}$

H/w: Prove that d is convex if c is a convex function
and if C is a convex set

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 $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x})||$, where, $P_C(\mathbf{x}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \dfrac{\mathbf{x} - P_C(\mathbf{x})}{\|\mathbf{x} - P_C(\mathbf{x})\|}$ ....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 $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x})||$, where, $P_C(\mathbf{x}) = \text{argmin } d(\mathbf{x}, \mathcal{C})$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function and $\nabla d(\mathbf{x}, \mathcal{C}) = \frac{\mathbf{x} - P_C(\mathbf{x})}{||\mathbf{x} - P_C(\mathbf{x})||}$ ....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: $prox_c(\mathbf{x}) = \underset{y}{\text{argmin }} PROX_c(\mathbf{x})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||$ The special case is when

c(x) is the indicator function over C

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 $d(\mathbf{x},\mathcal{C}) = \inf_{y\in\mathcal{C}} ||\mathbf{x}-\mathbf{y}|| = ||\mathbf{x} - P_C(\mathbf{x})||$, where, $P_C(\mathbf{x}) = \text{argmin } d(\mathbf{x},\mathcal{C})$. Then $d(\mathbf{x},\mathcal{C})$ is a convex function and $\nabla d(\mathbf{x},\mathcal{C}) = \dfrac{\mathbf{x}-P_C(\mathbf{x})}{\|\mathbf{x}-P_C(\mathbf{x})\|}$ ....The point of intersection of convex sets $C_1,\ C_2,\ldots C_m$ by minimizing... (Subgradients and Alternating Projections)
  - $\operatorname*{argmin}_{y\in\mathcal{C}} d(\mathbf{x},\mathcal{C})$ is a special case of the proximity operator: $prox_c(\mathbf{x}) = \operatorname*{argmin}_{y} PROX_c(\mathbf{x})$ of a convex function $c(\mathbf{x})$. Here, $PROX_c(\mathbf{x}) = c(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||$ 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 $\partial PROX_c(\mathbf{x}) = \partial c(\mathbf{y}) + \mathbf{y} - \mathbf{x}$ which can now be obtained for the special case $c(\mathbf{y}) = I_C(\mathbf{y})$.

Proximal will be done in details later