Some norms
(cont’d from last time)
Vector norms: recap

Example. The Euclidean or $\ell_2$-norm is $\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}$

Example. Let $p \geq 1$; $\ell_p$-norm is $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$

Exercise: Verify that $\|x\|_p$ is indeed a norm.

Example. ($\ell_\infty$-norm): $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Example. (Frobenius-norm): Let $A \in \mathbb{C}^{m \times n}$. The Frobenius norm of $A$ is $\|A\|_F := \sqrt{\sum_{ij} |a_{ij}|^2}$; that is, $\|A\|_F = \sqrt{\text{Tr}(A^*A)}$. 

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (2/18/21; Lecture 2)
Important example: Distance function

**Claim.** Let $\mathcal{Y}$ be a convex set. Let $x \in \mathbb{R}^d$ be some point. The distance of $x$ to the set $\mathcal{Y}$ is defined as

$$
\text{dist}(x, \mathcal{Y}) := \inf_{y \in \mathcal{Y}} \|x - y\|.
$$

**Proof.** Observe that $\|x - y\|$ is jointly convex in $(x, y)$ (Why?). Thus, the function $\text{dist}(x, \mathcal{Y})$ is a convex function of $x$ using the partial minimization rule.
Matrix Norms: induced norm

Let $A \in \mathbb{R}^{m \times n}$, and let $\| \cdot \|$ be any vector norm. We define an induced matrix norm as

$$
\|A\| := \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}.
$$
Matrix Norms: induced norm

Let \( A \in \mathbb{R}^{m \times n} \), and let \( \| \cdot \| \) be any vector norm. We define an 
induced matrix norm as

\[
\|A\| := \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}.
\]

Verify it is a norm

- Clearly, \( \|A\| = 0 \) iff \( A = 0 \) (definiteness)
- \( \|\alpha A\| = |\alpha| \|A\| \) (homogeneity)
- \( \|A + B\| = \sup \frac{\|(A+B)x\|}{\|x\|} \leq \sup \frac{\|Ax\|+\|Bx\|}{\|x\|} \leq \|A\| + \|B\| \).
Example. Let $A$ be any matrix. Its **operator norm** is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of $A$. 
Example. Let $A$ be any matrix. Its operator norm is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\text{max}}(A)$, where $\sigma_{\text{max}}$ is the largest singular value of $A$.

- **Warning!** Generally, largest eigenvalue **not** a norm!
Example. Let $A$ be any matrix. Its **operator norm** is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\text{max}}(A)$, where $\sigma_{\text{max}}$ is the largest singular value of $A$.

- **Warning!** Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
Example. Let $A$ be any matrix. Its **operator norm** is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\text{max}}(A)$, where $\sigma_{\text{max}}$ is the largest singular value of $A$.

- **Warning!** Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_{\infty}$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \not\in \{1, 2, \infty\}$
Example. Let $A$ be any matrix. Its **operator norm** is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\text{max}}(A)$, where $\sigma_{\text{max}}$ is the largest singular value of $A$.

- **Warning!** Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \not\in \{1, 2, \infty\}$
- **Schatten $p$-norm:** $\ell_p$-norm of vector of singular values.
**Example.** Let $A$ be any matrix. Its **operator norm** is

$$
\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.
$$

It can be shown that $\|A\|_2 = \sigma_{\text{max}}(A)$, where $\sigma_{\text{max}}$ is the largest singular value of $A$.

- **Warning!** Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \not\in \{1, 2, \infty\}$
- **Schatten $p$-norm:** $\ell_p$-norm of vector of singular values.
- **Exercise:** Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ be singular values of a matrix $A \in \mathbb{R}^{m \times n}$. Prove that

$$
\|A\|_{(k)} := \sum_{i=1}^{k} \sigma_i(A),
$$

is a norm; $1 \leq k \leq n$. 
Def. Support function: \( \sigma_C(x) = \sup_{z \in C} z^T x \)
Support function and dual norms

**Def. Support function:** $\sigma_{C}(x) = \sup_{z \in C} z^T x$

Support function for the unit norm ball is called: dual norm.

**Def.** Let $\| \cdot \|$ be a norm on $\mathbb{R}^d$. Its **dual norm** is

$$\| u \|_* := \sup \{ u^T x | \| x \| \leq 1 \} = \sigma_{\| x \| \leq 1}(u).$$

**Exercise:** Verify that $\| u \|_*$ is a norm.
Support function and dual norms

Def. Support function: \( \sigma_C(x) = \sup_{z \in C} z^T x \)

Support function for the unit norm ball is called: dual norm.

Def. Let \( \| \cdot \| \) be a norm on \( \mathbb{R}^d \). Its dual norm is

\[
\| u \|_\ast := \sup \{ u^T x \mid \| x \| \leq 1 \} = \sigma_{\| x \| \leq 1}(u).
\]

Exercise: Verify that \( \| u \|_\ast \) is a norm.

Exercise: Let \( 1/p + 1/q = 1 \), where \( p, q \geq 1 \). Show that \( \| \cdot \|_q \) is dual to \( \| \cdot \|_p \). In particular, the \( \ell_2 \)-norm is self-dual.

Exercise. Verify the generalized Hölder inequality \( u^T x \leq \| u \| \| x \|_\ast \) using the definition of dual norms.
**Def.** Let $K, L \subseteq \mathbb{R}^d$ be two sets. The **Hausdorff distance** between them is defined as

$$d_H(K, L) := \inf \{ \lambda \geq 0 \mid K \subseteq L + \lambda B(0, 1), L \subseteq K + \lambda B(0, 1) \}.$$ 

(See e.g., [https://en.wikipedia.org/wiki/Hausdorff_distance](https://en.wikipedia.org/wiki/Hausdorff_distance))
Support functions and Hausdorff distance

**Def.** Let $K, L \subseteq \mathbb{R}^d$ be two sets. The **Hausdorff distance** between them is defined as

$$d_H(K, L) := \inf \{ \lambda \geq 0 \mid K \subseteq L + \lambda B(0, 1), L \subseteq K + \lambda B(0, 1) \}.$$  

(See e.g., [https://en.wikipedia.org/wiki/Hausdorff_distance](https://en.wikipedia.org/wiki/Hausdorff_distance))

**Lemma** Let $K, L$ be convex bodies in $\mathbb{R}^d$. Then,

$$d_H(K, L) = \sup_{\|u\|_2 \leq 1} |\sigma_K(u) - \sigma_L(u)|.$$  

**Explore.** Support functions are important in the subject of **convex geometry**; read up on them and explore a bit!
Fenchel conjugates

Convex analysis analog of Fourier transforms:

Def. Fenchel conjugate: \( f^*(y) := \sup_{x \in \text{dom} f} \langle x, y \rangle - f(x) \)
Convex analysis analog of Fourier transforms:

**Def. Fenchel conjugate:** \( f^*(y) := \sup_{x \in \text{dom}} f\langle x, y \rangle - f(x) \)

**Observe:** \( f^* \) is convex, even if \( f \) is not. If \( f \) differentiable, then \( f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle - f(x) \) (Fenchel-Legendre transform).
Fenchel conjugates

Convex analysis analog of Fourier transforms:

**Def. Fenchel conjugate:** \( f^*(y) := \sup_{x \in \text{dom} f} \langle x, y \rangle - f(x) \)

**Observe:** \( f^* \) is convex, even if \( f \) is not. If \( f \) differentiable, then \( f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle - f(x) \) (Fenchel-Legendre transform).

**Fenchel-Young inequality:** \( f^*(u) + f(x) \geq \langle u, x \rangle \)
Fenchel conjugates

Convex analysis analog of Fourier transforms:

**Def. Fenchel conjugate:**
\[ f^*(y) := \sup_{x \in \text{dom} f} \langle x, y \rangle - f(x) \]

**Observe:** $f^*$ is convex, even if $f$ is not. If $f$ differentiable, then
\[ f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle - f(x) \] (Fenchel-Legendre transform).

**Fenchel-Young inequality:**
\[ f^*(u) + f(x) \geq \langle u, x \rangle \]

Fenchel transforms satisfy the beautiful *duality* property:

**Theorem.** Let $f$ be a closed convex function (i.e., epi $f = \{(x, t) \mid f(x) \leq t\}$ is a closed convex set; equivalently, $f$ is lower semi-continuous). Then, $f^{**} = f$. 
Fenchel conjugates

Convex analysis analog of Fourier transforms:

**Def. Fenchel conjugate:** \( f^*(y) := \sup_{x \in \text{dom} f} \langle x, y \rangle - f(x) \)

**Observe:** \( f^* \) is convex, even if \( f \) is not. If \( f \) differentiable, then \( f^*(\nabla f(x)) = \langle x, \nabla f(x) \rangle - f(x) \) (Fenchel-Legendre transform).

**Fenchel-Young inequality:** \( f^*(u) + f(x) \geq \langle u, x \rangle \)

Fenchel transforms satisfy the beautiful *duality* property:

**Theorem.** Let \( f \) be a closed convex function (i.e., \( \text{epi } f = \{ (x,t) \mid f(x) \leq t \} \) is a closed convex set; equivalently, \( f \) is lower semi-continuous). Then, \( f^{**} = f \).

**Exercise:** Show that \( f^* = f \iff f = \frac{1}{2} \| \cdot \|_2^2 \).
Example. $f(x) = ax + b$; then,

$$f^*(z) = \sup_x z x - (ax + b)$$
Example. \( f(x) = ax + b \); then,

\[
\begin{align*}
    f^*(z) &= \sup_x zx - (ax + b) \\
    &= \infty, \quad \text{if } (z - a) \neq 0.
\end{align*}
\]
**Example.** \( f(x) = ax + b; \) then,

\[
f^*(z) = \sup_x zx - (ax + b) = \infty, \quad \text{if} \ (z - a) \neq 0.
\]

Thus, \( \text{dom} f^* = \{a\} \), and \( f^*(a) = -b \).
Example. $f(x) = ax + b$; then,

$$f^*(z) = \sup_x z x - (ax + b)$$

$$= \infty, \quad \text{if } (z - a) \neq 0.$$ 

Thus, $\text{dom} f^* = \{a\}$, and $f^*(a) = -b$.

Example. Let $a \geq 0$, and set $f(x) = -\sqrt{a^2 - x^2}$ if $|x| \leq a$, and $+\infty$ otherwise. Then, $f^*(z) = a\sqrt{1 + z^2}$. 
Fenchel conjugate – examples

**Example.** \( f(x) = ax + b; \) then,
\[
f^*(z) = \sup_x zx - (ax + b)
\]
\[
= \infty, \quad \text{if } (z - a) \neq 0.
\]

Thus, \( \text{dom } f^* = \{a\}, \) and \( f^*(a) = -b. \)

**Example.** Let \( a \geq 0, \) and set \( f(x) = -\sqrt{a^2 - x^2} \) if \( |x| \leq a, \) and \( +\infty \) otherwise. Then, \( f^*(z) = a\sqrt{1 + z^2}. \)

**Example.** \( f(x) = \frac{1}{2} x^T Ax, \) where \( A \succ 0. \) Then, \( f^*(z) = \frac{1}{2} z^T A^{-1} z. \)
Example. \( f(x) = ax + b; \) then,

\[
    f^*(z) = \sup_x z x - (ax + b) \\
    = \infty, \quad \text{if } (z - a) \neq 0.
\]

Thus, \( \text{dom} f^* = \{a\} \), and \( f^*(a) = -b \).

Example. Let \( a \geq 0 \), and set \( f(x) = -\sqrt{a^2 - x^2} \) if \( |x| \leq a \), and \( +\infty \) otherwise. Then, \( f^*(z) = a\sqrt{1 + z^2} \).

Example. \( f(x) = \frac{1}{2} x^T A x, \) where \( A \succ 0 \). Then, \( f^*(z) = \frac{1}{2} z^T A^{-1} z \).

Exercise: If \( f(x) = \max(0, 1 - x) \), then \( \text{dom} f^* \) is \([-1, 0]\), and within this domain, \( f^*(z) = z \).
Fenchel conjugate of norms

Recall: Dual norm

\[ \|u\|_* := \sup \{ u^T x \mid \|x\| \leq 1 \} . \]
Fenchel conjugate of norms

Recall: Dual norm

\[ \|u\|_* := \sup \{ u^T x \mid \|x\| \leq 1 \}. \]

**Example.** Let \( f(x) = \|x\| \). We have \( f^*(z) = \delta_{\|\cdot\|_* \leq 1}(z) \). Thus, conjugate of a norm is the *indicator of unit dual norm ball*. 
Fenchel conjugate of norms

Recall: Dual norm

\[ \|u\|_* := \sup \{ u^T x | \|x\| \leq 1 \} \].

Example. Let \( f(x) = \|x\| \). We have \( f^*(z) = \delta_{\|\cdot\|_* \leq 1}(z) \). Thus, conjugate of a norm is the **indicator of unit dual norm ball**.

Proof.

- Consider two cases: (i) \( \|z\|_* > 1 \); (ii) \( \|z\|_* \leq 1 \)
- (i): by def. of dual norm there is a \( u \) s.t. \( \|u\| \leq 1 \) and \( z^T u > 1 \)
- \( f^*(z) = \sup_x x^T z - f(x) \). **Rewrite** \( x = \alpha u \), and let \( \alpha \to \infty \)
- Then, \( z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha (z^T u - \|u\|) \); \( \to \infty \)
- Case (ii): Since \( z^T x \leq \|x\| \|z\|_* \), \( x^T z - \|x\| \leq \|x\| (\|z\|_* - 1) \leq 0 \).
- \( x = 0 \) maximizes \( \|x\| (\|z\|_* - 1) \), hence \( f(z) = 0 \).
- Thus, \( f^*(z) = +\infty \) if (i), and 0 if (ii), completing the proof.
In Fourier analysis, we discover that convolution can be described via the product of Fourier transforms.

In convex analysis, the counterpart is infimal convolution $(f \Box g)(x) := \inf_{y \in X} f(y) + g(x - y)$, where both $f$ and $g$ are (suitable) convex functions.

Then, under appropriate hypotheses one has $(f \Box g)^* = f^* + g^*$, and $(f + g)^* = f^* \Box g^*$.
Fenchel conjugates – analogies

- In Fourier analysis, we discover that *convolution* can be described via the product of Fourier transforms.
- In convex analysis, the counterpart is *infimal convolution*

\[(f \square g)(x) := \inf_{y \in X} f(y) + g(x - y),\]

where both \(f\) and \(g\) are (suitable) convex functions.
In Fourier analysis, we discover that *convolution* can be described via the product of Fourier transforms.

In convex analysis, the counterpart is *infimal convolution*

\[
(f \boxplus g)(x) := \inf_{y \in X} f(y) + g(x - y),
\]

where both \(f\) and \(g\) are (suitable) convex functions.

Then, under appropriate hypotheses one has

\[
(f \boxplus g)^* = f^* + g^*, \quad \text{and} \quad (f + g)^* = f^* \boxplus g^*.
\]
In Fourier analysis, we discover that *convolution* can be described via the product of Fourier transforms.

In convex analysis, the counterpart is *infimal convolution*

\[(f \square g)(x) := \inf_{y \in X} f(y) + g(x - y),\]

where both \(f\) and \(g\) are (suitable) convex functions.

Then, under appropriate hypotheses one has

\[(f \square g)^* = f^* + g^*, \quad \text{and} \quad (f + g)^* = f^* \square g^*.\]

**Challenge.** Recall: \(f(x) = \frac{1}{2}x^T Ax (A \succ 0)\) then \(f^*(z) = \frac{1}{2}z^T A^{-1}z\). Let \(f_i(x) := x^T A_i x\) for \(A_i \succ 0\) and \(1 \leq i \leq n\). Consider,

\[F(z) := \sum_i f_i^*(z) - \sum_{i<j} (f_i + f_j)^*(z) + \cdots + (-1)^{n+1}(f_1 + \cdots + f_n)^*(z).\]

Prove or disprove that \(F\) is convex.
Fenchel conjugates are special

Let $\Gamma_0(\mathbb{R}^d)$ denote class of closed, convex functions on $\mathbb{R}^d$. The (Legendre)-Fenchel transform of $f \in \Gamma_0$ is defined as

$$\mathcal{L} : f \mapsto \sup_{y} \langle \cdot, y \rangle - f(y)$$

(so that $(\mathcal{L}f)(x) = f^*(x)$).

Theorem.

Let $T$ be a transform that maps $\Gamma_0(\mathbb{R}^d)$ to $\Gamma_0(\mathbb{R}^d)$ and satisfies: (i) $T(Tf) = f$ (closure); and (ii) $f \leq g \Rightarrow Tf \geq Tg$.

Then, $T$ must "essentially" be the Fenchel transform. More precisely, there exists $c \in \mathbb{R}$, $v \in \mathbb{R}^d$ and $B \in \text{GL}_n(\mathbb{R})$ such that

$$(Tf)(x) = (\mathcal{L}f)(Bx + v) + \langle v, x \rangle + c$$

Explore:

Study other classes instead of $\Gamma_0(\mathbb{R}^d)$ for which similar theorems can be proved.
Fenchel conjugates are special

Let $\Gamma_0(\mathbb{R}^d)$ denote class of closed, convex functions on $\mathbb{R}^d$. The (Legendre)-Fenchel transform of $f \in \Gamma_0$ is defined as

$$\mathcal{L} : f \mapsto \sup_y \langle \cdot, y \rangle - f(y)$$

(so that $(\mathcal{L}f)(x) = f^*(x)$).

**Theorem.** Let $\mathcal{T}$ be a transform that maps $\Gamma_0 \rightarrow \Gamma_0$ and satisfies: (i) $\mathcal{T}(\mathcal{T}f) = f$ (closure); and (ii) $f \leq g \implies \mathcal{T}f \geq \mathcal{T}g$.

Then, $\mathcal{T}$ must “essentially” be the Fenchel transform. More precisely, there exists $c \in \mathbb{R}, v \in \mathbb{R}^d$ and $B \in GL_n(\mathbb{R})$ such that

$$(\mathcal{T}f)(x) = (\mathcal{L}f)(Bx + v) + \langle v, x \rangle + c$$

**Explore:** Study other classes instead of $\Gamma_0(\mathbb{R}^d)$ for which similar theorems can be proved.
Subdifferentials

DO: (Read S. Boyd’s EE364B notes)
First order global underestimator

\[ f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \]
First order global underestimator

\[ f(x) \geq f(y) + \langle g, x - y \rangle \]
Subgradients

$g_1, g_2, g_3$ are subgradients at $y$
Subgradients – basic facts

- $f$ is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at $y$
- A vector $g$ is a subgradient at a point $y$ if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Often **one** subgradient costs approx as much as $f(x)$
Subgradients – basic facts

- $f$ is convex, differentiable: $\nabla f(y)$ the \textit{unique} subgradient at $y$
- A vector $g$ is a subgradient at a point $y$ if and only if $f(y) + \langle g, x - y \rangle$ is \textit{globally} smaller than $f(x)$.
- Often \textit{one} subgradient costs approx as much as $f(x)$
- Determining \textit{all} subgradients at a given point — \textit{difficult}.
- Subgradient calculus: great achievement in convex analysis
Subgradients – basic facts

- $f$ is convex, differentiable: $\nabla f(y)$ the unique subgradient at $y$
- A vector $g$ is a subgradient at a point $y$ if and only if $f(y) + \langle g, x - y \rangle$ is globally smaller than $f(x)$.
- Often one subgradient costs approx as much as $f(x)$
- Determining all subgradients at a given point — difficult.
- Subgradient calculus: great achievement in convex analysis
- Without convexity, things become wild (e.g., chain rule fails!)
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]
Subgradients – example

\( f(x) := \max(f_1(x), f_2(x)) \); both \( f_1, f_2 \) convex, differentiable
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)) ; \text{both } f_1, f_2 \text{ convex, differentiable} \]

\[ f_1(x) \]
\[ f_2(x) \]
\[ f(x) \]

\[ \star f_1(x) > f_2(x) : \text{unique subgradient of } f \text{ is } f'_1(x) \]
$$f(x) := \max(f_1(x), f_2(x));\ \text{both } f_1, f_2 \text{ convex, differentiable}$$

\[ f(x) \]
\[ f_1(x) \]
\[ f_2(x) \]

$\blackstar f_1(x) > f_2(x)$: unique subgradient of $f$ is $f_1'(x)$

$\blackstar f_1(x) < f_2(x)$: unique subgradient of $f$ is $f_2'(x)$
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]

\[ f_1(x) \]
\[ f_2(x) \]
\[ f(x) \]

\[ y \]

\[ f_1(x) > f_2(x): \text{ unique subgradient of } f \text{ is } f_1'(x) \]

\[ f_1(x) < f_2(x): \text{ unique subgradient of } f \text{ is } f_2'(x) \]

\[ f_1(y) = f_2(y): \text{ subgradients, the segment } [f_1'(y), f_2'(y)] \]
\[ (\text{imagine all supporting lines turning about point } y) \]
Def. A vector $g \in \mathbb{R}^n$ is called a subgradient at a point $y$, if for all $x \in \text{dom} f$, it holds that
\[ f(x) \geq f(y) + \langle g, x - y \rangle \]

Def. The set of all subgradients at $y$ denoted by $\partial f(y)$. This set is called subdifferential of $f$ at $y$
**Subgradients and the Subdifferential (Set)**

**Def.** A vector \( g \in \mathbb{R}^n \) is called a **subgradient** at a point \( y \), if for all \( x \in \text{dom} \, f \), it holds that

\[
f(x) \geq f(y) + \langle g, x - y \rangle
\]

**Def.** The set of all subgradients at \( y \) denoted by \( \partial f(y) \). This set is called **subdifferential** of \( f \) at \( y \)

If \( f \) is convex, \( \partial f(x) \) is nice:

♣ If \( x \in \text{relative interior of} \, \text{dom} \, f \), then \( \partial f(x) \) nonempty
Def. A vector $g \in \mathbb{R}^n$ is called a subgradient at a point $y$, if for all $x \in \text{dom } f$, it holds that

$$f(x) \geq f(y) + \langle g, x - y \rangle$$

Def. The set of all subgradients at $y$ denoted by $\partial f(y)$. This set is called subdifferential of $f$ at $y$.

If $f$ is convex, $\partial f(x)$ is nice:

♣ If $x \in$ relative interior of $\text{dom } f$, then $\partial f(x)$ nonempty

♣ If $f$ differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$
**Subgradients and the Subdifferential (Set)**

**Def.** A vector \( g \in \mathbb{R}^n \) is called a **subgradient** at a point \( y \), if for all \( x \in \text{dom} \ f \), it holds that

\[
f(x) \geq f(y) + \langle g, x - y \rangle
\]

**Def.** The set of all subgradients at \( y \) denoted by \( \partial f(y) \). This set is called **subdifferential** of \( f \) at \( y \)

If \( f \) is convex, \( \partial f(x) \) is nice:

- ✦ If \( x \in \text{relative interior of} \ \text{dom} \ f \), then \( \partial f(x) \) nonempty
- ✦ If \( f \) differentiable at \( x \), then \( \partial f(x) = \{ \nabla f(x) \} \)
- ♦ If \( \partial f(x) = \{ g \} \), then \( f \) is differentiable and \( g = \nabla f(x) \)
Subdifferential – example

\[ f(x) = |x| \]
Subdifferential – example

\[ f(x) = |x| \]
Subdifferential – example

\[ f(x) = |x| \]

\[ \partial f(x) = \begin{cases} 
-1 & \text{if } x < 0, \\
+1 & \text{if } x > 0, \\
[-1, 1] & \text{if } x = 0.
\end{cases} \]
More examples

**Example.** $f(x) = \|x\|_2$. Then,

$$
\partial f(x) := \begin{cases}
\|x\|_2^{-1} x & x \neq 0, \\
\{z \mid \|z\|_2 \leq 1\} & x = 0.
\end{cases}
$$
More examples

Example. $f(x) = \|x\|_2$. Then,

\[
\partial f(x) := \begin{cases} 
\|x\|_2^{-1}x & x \neq 0, \\
\{z \mid \|z\|_2 \leq 1\} & x = 0.
\end{cases}
\]

Proof.

\[
\begin{align*}
\|z\|_2 & \geq \|x\|_2 + \langle g, z - x \rangle \\
\|z\|_2 & \geq \langle g, z \rangle \\
\implies & \|g\|_2 \leq 1.
\end{align*}
\]
Calculus rules
If $f$ and $k$ are differentiable, we know that

- **Addition:** $\nabla (f + k)(x) = \nabla f(x) + \nabla k(x)$
- **Scaling:** $\nabla (\alpha f(x)) = \alpha \nabla f(x)$
Recall basic calculus

If $f$ and $k$ are differentiable, we know that

- **Addition**: $\nabla (f + k)(x) = \nabla f(x) + \nabla k(x)$
- **Scaling**: $\nabla (\alpha f(x)) = \alpha \nabla f(x)$

**Chain rule**

If $f : \mathbb{R}^n \to \mathbb{R}^m$, and $k : \mathbb{R}^m \to \mathbb{R}^p$. Let $h : \mathbb{R}^n \to \mathbb{R}^p$ be the composition $h(x) = (k \circ f)(x) = k(f(x))$. Then,

$$Dh(x) = Dk(f(x))Df(x).$$
Recall basic calculus

If $f$ and $k$ are differentiable, we know that

- **Addition:** $\nabla (f + k)(x) = \nabla f(x) + \nabla k(x)$
- **Scaling:** $\nabla (\alpha f(x)) = \alpha \nabla f(x)$

**Chain rule**

If $f : \mathbb{R}^n \to \mathbb{R}^m$, and $k : \mathbb{R}^m \to \mathbb{R}^p$. Let $h : \mathbb{R}^n \to \mathbb{R}^p$ be the composition $h(x) = (k \circ f)(x) = k(f(x))$. Then,

$$Dh(x) = Dk(f(x))Df(x).$$

**Example.** If $f : \mathbb{R}^n \to \mathbb{R}$ and $k : \mathbb{R} \to \mathbb{R}$, then using the fact that $\nabla h(x) = [Dh(x)]^T$, we obtain

$$\nabla h(x) = k'(f(x))\nabla f(x).$$
Finding one subgradient within $\partial f(x)$
Subgradient calculus

♠ Finding one subgradient within $\partial f(x)$
♠ Determining entire subdifferential $\partial f(x)$ at a point $x$
Subgradient calculus

♠ Finding one subgradient within $\partial f(x)$
♠ Determining entire subdifferential $\partial f(x)$ at a point $x$
♠ Do we have the chain rule?
Subgradient calculus

♠ Finding one subgradient within $\partial f(x)$
♠ Determining entire subdifferential $\partial f(x)$ at a point $x$
♠ Do we have the chain rule?
♠ Usually not easy!
If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$
If \( f \) is differentiable, \( \partial f(x) = \{ \nabla f(x) \} \)

Scaling \( \alpha > 0, \partial(\alpha f)(x) = \alpha \partial f(x) = \{ \alpha g \mid g \in \partial f(x) \} \)
Subgradient calculus

If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$

Scaling $\alpha > 0$, $\partial (\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$

Addition*: $\partial (f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
Subgradient calculus

- If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$

- **Scaling** $\alpha > 0$, $\partial (\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$

- **Addition***: $\partial (f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)

- **Chain rule***: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,

$$\partial h(x) = A^T \partial f(Ax + b).$$
Subgradient calculus

- If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- **Scaling** $\alpha > 0$, $\partial (\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- **Addition***: $\partial (f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- **Chain rule***: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,
  \[
  \partial h(x) = A^T \partial f(Ax + b).
  \]
- **Chain rule***: $h(x) = f \circ k$, where $k : X \to Y$ is diff.
  \[
  \partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))
  \]
Subgradient calculus

- If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- **Scaling** $\alpha > 0$, $\partial (\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- **Addition***: $\partial (f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- **Chain rule***: Let $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,
  $$\partial h(x) = A^T \partial f(Ax + b).$$
- **Chain rule***: $h(x) = f \circ k$, where $k : X \to Y$ is diff.
  $$\partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))$$
- **Max function***: If $f(x) := \max_{1 \leq i \leq m} f_i(x)$, then
  $$\partial f(x) = \text{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$
  convex hull over subdifferentials of “active” functions at $x$
Subgradient calculus

- If \( f \) is differentiable, \( \partial f(x) = \{ \nabla f(x) \} \)
- Scaling \( \alpha > 0 \), \( \partial (\alpha f)(x) = \alpha \partial f(x) = \{ \alpha g \mid g \in \partial f(x) \} \)
- Addition*: \( \partial (f + k)(x) = \partial f(x) + \partial k(x) \) (set addition)
- Chain rule*: Let \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, f : \mathbb{R}^m \to \mathbb{R} \), and \( h : \mathbb{R}^n \to \mathbb{R} \) be given by \( h(x) = f(Ax + b) \). Then,
  \[
  \partial h(x) = A^T \partial f(Ax + b).
  \]
- Chain rule*: \( h(x) = f \circ k \), where \( k : X \to Y \) is diff.
  \[
  \partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))
  \]
- Max function*: If \( f(x) := \max_{1 \leq i \leq m} f_i(x) \), then
  \[
  \partial f(x) = \text{conv} \bigcup \{ \partial f_i(x) \mid f_i(x) = f(x) \},
  \]
  convex hull over subdifferentials of “active” functions at \( x \)
- Conjugation: \( z \in \partial f(x) \) if and only if \( x \in \partial f^*(z) \)
Failure of addition rule

It can happen that $\partial (f_1 + f_2) \neq \partial f_1 + \partial f_2$
Failure of addition rule

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

**Example.** Define $f_1$ and $f_2$ by

$$f_1(x) := \begin{cases} 
-2\sqrt{x} & \text{if } x \geq 0, \\
+\infty & \text{if } x < 0,
\end{cases} \quad \text{and} \quad f_2(x) := \begin{cases} 
+\infty & \text{if } x > 0, \\
-2\sqrt{-x} & \text{if } x \leq 0.
\end{cases}$$

Then, $f = f_1 + f_2 = 1_0$, whereby $\partial f(0) = \mathbb{R}$

But $\partial f_1(0) = \partial f_2(0) = \emptyset.$
Failure of addition rule

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example. Define $f_1$ and $f_2$ by

$$f_1(x) := \begin{cases} 
-2\sqrt{x} & \text{if } x \geq 0, \\
+\infty & \text{if } x < 0,
\end{cases}$$

and $f_2(x) := \begin{cases} 
+\infty & \text{if } x > 0, \\
-2\sqrt{-x} & \text{if } x \leq 0.
\end{cases}$

Then, $f = f_1 + f_2 = 1_0$, whereby $\partial f(0) = \mathbb{R}$
But $\partial f_1(0) = \partial f_2(0) = \emptyset$.

However, $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ always holds.

Exercise: Prove the above statement.
Example. $f(x) = \|x\|_\infty$. Then,

$$\partial f(0) = \text{conv} \{\pm e_1, \ldots, \pm e_n\},$$

where $e_i$ is $i$-th canonical basis vector.
Subdifferential: two examples

Example. $f(x) = \|x\|_\infty$. Then,

$$\partial f(0) = \text{conv} \{ \pm e_1, \ldots, \pm e_n \},$$

where $e_i$ is $i$-th canonical basis vector

To prove, notice that $f(x) = \max_{1 \leq i \leq n} \{|e_i^T x|\}$; apply max rule.
**Example.** $f(x) = \|x\|_\infty$. Then,

$$\partial f(0) = \text{conv} \left\{ \pm e_1, \ldots, \pm e_n \right\},$$

where $e_i$ is $i$-th canonical basis vector

To prove, notice that $f(x) = \max_{1 \leq i \leq n} \{ |e_i^T x| \}$; apply max rule.

**Example.** Let $f_1, f_2, \ldots, f_m$ be differentiable and convex. Let

$$f(x) := \max(f_1(x), \ldots, f_m(x))$$

$$\partial f(x) = \text{co} \left\{ \nabla f_i(x) \mid f_i(x) = f(x) \right\}$$
Computing subgradients
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in Y} h(x, y) \]

Getting \( \partial f(x) \) is complicated!
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting \( \partial f(x) \) is complicated!

Simple way to obtain some \( g \in \partial f(x) \):
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting \( \partial f(x) \) is complicated!

Simple way to obtain some \( g \in \partial f(x) \):

- Pick any \( y^* \) for which \( h(x, y^*) = f(x) \)
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting \( \partial f(x) \) is complicated!

Simple way to obtain some \( g \in \partial f(x) \):

- Pick any \( y^* \) for which \( h(x, y^*) = f(x) \)
- Pick any subgradient \( g \in \partial h(x, y^*) \)
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting $\partial f(x)$ is complicated!

Simple way to obtain some $g \in \partial f(x)$:

- Pick any $y^*$ for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$
- This $g \in \partial f(x)$
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting \( \partial f(x) \) is complicated!

Simple way to obtain some \( g \in \partial f(x) \):

- Pick any \( y^* \) for which \( h(x, y^*) = f(x) \)
- Pick any subgradient \( g \in \partial h(x, y^*) \)
- This \( g \in \partial f(x) \)

\[
\begin{align*}
  h(z, y^*) & \geq h(x, y^*) + g^T (z - x) \\
  h(z, y^*) & \geq f(x) + g^T (z - x)
\end{align*}
\]
Subgradient for pointwise sup

\[ f(x) := \sup_{y \in \mathcal{Y}} h(x, y) \]

Getting \( \partial f(x) \) is complicated!

Simple way to obtain some \( g \in \partial f(x) \):

\> Pick any \( y^* \) for which \( h(x, y^*) = f(x) \)

\> Pick any subgradient \( g \in \partial h(x, y^*) \)

\> This \( g \in \partial f(x) \)

\[
\begin{align*}
  h(z, y^*) & \geq h(x, y^*) + g^T(z - x) \\
  h(z, y^*) & \geq f(x) + g^T(z - x) \\
  f(z) & \geq h(z, y^*) \quad \text{(because of sup)} \\
  f(z) & \geq f(x) + g^T(z - x).
\end{align*}
\]
Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).$$

This $f$ a max (in fact, over a finite number of terms)
Example

Suppose \( a_i \in \mathbb{R}^n \) and \( b_i \in \mathbb{R} \). And

\[
f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).
\]

This \( f \) a max (in fact, over a finite number of terms)

| ➤ Suppose \( f(x) = a_k^T x + b_k \) for some index \( k \) |
Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).$$

This $f$ a max (in fact, over a finite number of terms)

- Suppose $f(x) = a_k^T x + b_k$ for some index $k$
- Here $f(x; y) = f_k(x) = a_k^T x + b_k$, and $\partial f_k(x) = \{\nabla f_k(x)\}$
Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).$$

This $f$ a max (in fact, over a finite number of terms)

- Suppose $f(x) = a_k^T x + b_k$ for some index $k$
- Here $f(x; y) = f_k(x) = a_k^T x + b_k$, and $\partial f_k(x) = \{\nabla f_k(x)\}$
- Hence, $a_k \in \partial f(x)$ works!
Suppose $f = \mathbb{E}f(x, u)$, where $f$ is convex in $x$ for each $u$ (an r.v.)

$$f(x) := \int f(x, u)p(u)du$$
Subgradient of expectation

Suppose $f = \mathbf{E} f(x, u)$, where $f$ is convex in $x$ for each $u$ (an r.v.)

$$f(x) := \int f(x, u)p(u)du$$

- For each $u$ choose any $g(x, u) \in \partial_x f(x, u)$
Subgradient of expectation

Suppose $f = \mathbb{E} f(x, u)$, where $f$ is convex in $x$ for each $u$ (an r.v.)

$$f(x) := \int f(x, u)p(u)du$$

- For each $u$ choose any $g(x, u) \in \partial_x f(x, u)$
- Then, $g = \int g(x, u)p(u)du = \mathbb{E}g(x, u) \in \partial f(x)$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is convex and increasing; each $f_i$ is convex

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$
Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and increasing; each $f_i$ cvx.

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

We can find a vector $g \in \partial f(x)$ as follows:

\[ g_i \in \partial f_i(x) \]
\[ u \in \partial h(f_1(x), \ldots, f_n(x)) \]
\[ g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n; \]
\[ g \in \partial f(x) \]

Exercise: Verify $g \in \partial f(x)$ by showing $f(z) \geq f(x) + g^T(z - x)$.
Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and increasing; each $f_i$ cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

We can find a vector $g \in \partial f(x)$ as follows:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is convex and increasing; each $f_i$ is convex

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

We can find a vector $g \in \partial f(x)$ as follows:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is convex and increasing; each $f_i$ is convex.

Define $f(x) := h(f_1(x), f_2(x), \ldots, f_n(x))$.

We can find a vector $g \in \partial f(x)$ as follows:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$.
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$.
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$. 

Compare with $\nabla f(x) = J \nabla h(x)$, where $J$ is the gradient matrix of $f_i(x)$.
Suppose \( h : \mathbb{R}^n \rightarrow \mathbb{R} \) cvx and increasing; each \( f_i \) cvx

\[
f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).
\]

We can find a vector \( g \in \partial f(x) \) as follows:

- For \( i = 1 \) to \( n \), compute \( g_i \in \partial f_i(x) \)
- Compute \( u \in \partial h(f_1(x), \ldots, f_n(x)) \)
- Set \( g = u_1g_1 + u_2g_2 + \cdots + u_ng_n \); this \( g \in \partial f(x) \)
- Compare with \( \nabla f(x) = J\nabla h(x) \), where \( J \) matrix of \( \nabla f_i(x) \)
Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and increasing; each $f_i$ cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

We can find a vector $g \in \partial f(x)$ as follows:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$
- Compare with $\nabla f(x) = J \nabla h(x)$, where $J$ matrix of $\nabla f_i(x)$

**Exercise:** Verify $g \in \partial f(x)$ by showing $f(z) \geq f(x) + g^T(z - x)$