Subgradients

- subgradients
- strong and weak subgradient calculus
- optimality conditions via subgradients
- directional derivatives
Basic inequality

recall basic inequality for convex differentiable $f$:

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

• first-order approximation of $f$ at $x$ is global underestimator

• $(\nabla f(x), -1)$ supports $\text{epi} \ f$ at $(x, f(x))$

what if $f$ is not differentiable?
Subgradient of a function

$g$ is a subgradient of $f$ (not necessarily convex) at $x$ if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

$g_2, g_3$ are subgradients at $x_2$; $g_1$ is a subgradient at $x_1$
• $g$ is a subgradient of $f$ at $x$ iff $(g, -1)$ supports $\text{epi } f$ at $(x, f(x))$
• $g$ is a subgradient iff $f(x) + g^T(y - x)$ is a global (affine) underestimator of $f$
• if $f$ is convex and differentiable, $\nabla f(x)$ is a subgradient of $f$ at $x$

Subgradients come up in several contexts:
• algorithms for nondifferentiable convex optimization
• convex analysis, e.g., optimality conditions, duality for nondifferentiable problems

(if $f(y) \leq f(x) + g^T(y - x)$ for all $y$, then $g$ is a **supergradient**)
Example

\[ f = \max\{f_1, f_2\}, \text{ with } f_1, f_2 \text{ convex and differentiable} \]

- \( f_1(x_0) > f_2(x_0) \): unique subgradient \( g = \nabla f_1(x_0) \)
- \( f_2(x_0) > f_1(x_0) \): unique subgradient \( g = \nabla f_2(x_0) \)
- \( f_1(x_0) = f_2(x_0) \): subgradients form a line segment \([\nabla f_1(x_0), \nabla f_2(x_0)]\)
Subdifferential

- set of all subgradients of $f$ at $x$ is called the **subdifferential** of $f$ at $x$, denoted $\partial f(x)$

- $\partial f(x)$ is a closed convex set (can be empty)

If $f$ is convex,

- $\partial f(x)$ is nonempty, for $x \in \text{relint dom } f$
- $\partial f(x) = \{\nabla f(x)\}$, if $f$ is differentiable at $x$
- if $\partial f(x) = \{g\}$, then $f$ is differentiable at $x$ and $g = \nabla f(x)$
Example

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

righthand plot shows \( \bigcup \{(x, g) \mid x \in \mathbb{R}, \ g \in \partial f(x)\} \)
Subgradient calculus

- **weak subgradient calculus**: formulas for finding *one* subgradient $g \in \partial f(x)$

- **strong subgradient calculus**: formulas for finding the whole subdifferential $\partial f(x)$, *i.e.*, *all* subgradients of $f$ at $x$

- many algorithms for nondifferentiable convex optimization require only *one* subgradient at each step, so weak calculus suffices

- some algorithms, optimality conditions, etc., need whole subdifferential

- roughly speaking: if you can compute $f(x)$, you can usually compute a $g \in \partial f(x)$

- we’ll assume that $f$ is convex, and $x \in \text{relint dom } f$
Some basic rules

• $\partial f(x) = \{\nabla f(x)\}$ if $f$ is differentiable at $x$

• **scaling:** $\partial(\alpha f) = \alpha \partial f$ (if $\alpha > 0$)

• **addition:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$ (RHS is addition of point-to-set mappings)

• **affine transformation of variables:** if $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$

• **finite pointwise maximum:** if $f = \max_{i=1,...,m} f_i$, then

$$\partial f(x) = \text{Co} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

i.e., convex hull of union of subdifferentials of ‘active’ functions at $x$
\[ f(x) = \max\{f_1(x), \ldots, f_m(x)\}, \text{ with } f_1, \ldots, f_m \text{ differentiable} \]

\[ \partial f(x) = \text{Co}\{\nabla f_i(x) \mid f_i(x) = f(x)\} \]

**example:** \[ f(x) = \|x\|_1 = \max\{s^T x \mid s_i \in \{-1, 1\}\} \]

\[ \partial f(x) \text{ at } x = (0, 0) \quad \text{at } x = (1, 0) \quad \text{at } x = (1, 1) \]
Pointwise supremum

if \( f = \sup_{\alpha \in A} f_{\alpha} \),

\[
\text{cl Co} \bigcup \{ \partial f_{\beta}(x) \mid f_{\beta}(x) = f(x) \} \subseteq \partial f(x)
\]

(usually get equality, but requires some technical conditions to hold, e.g.,
\( A \) compact, \( f_{\alpha} \) cts in \( x \) and \( \alpha \))

roughly speaking, \( \partial f(x) \) is closure of convex hull of union of
subdifferentials of active functions
Weak rule for pointwise supremum

\[ f = \sup_{\alpha \in A} f_\alpha \]

- find any \( \beta \) for which \( f_\beta(x) = f(x) \) (assuming supremum is achieved)
- choose any \( g \in \partial f_\beta(x) \)
- then, \( g \in \partial f(x) \)
example

\[ f(x) = \lambda_{\text{max}}(A(x)) = \sup_{\|y\|_2 = 1} y^T A(x)y \]

where \( A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n, A_i \in S^k \)

- \( f \) is pointwise supremum of \( g_y(x) = y^T A(x)y \) over \( \|y\|_2 = 1 \)

- \( g_y \) is affine in \( x \), with \( \nabla g_y(x) = (y^T A_1 y, \ldots, y^T A_n y) \)

- hence, \( \partial f(x) \supseteq \text{Co} \{ \nabla g_y \mid A(x)y = \lambda_{\text{max}}(A(x))y, \|y\|_2 = 1 \} \)

  (in fact equality holds here)

to find one subgradient at \( x \), can choose any unit eigenvector \( y \) associated with \( \lambda_{\text{max}}(A(x)) \); then

\[ (y^T A_1 y, \ldots, y^T A_n y) \in \partial f(x) \]
Expectation

• $f(x) = \mathbb{E} f(x, \omega)$, with $f$ convex in $x$ for each $\omega$, $\omega$ a random variable
• for each $\omega$, choose any $g_\omega \in \partial f(x, \omega)$ (so $\omega \mapsto g_\omega$ is a function)
• then, $g = \mathbb{E} g_\omega \in \partial f(x)$

Monte Carlo method for (approximately) computing $f(x)$ and a $g \in \partial f(x)$:

• generate independent samples $\omega_1, \ldots, \omega_K$ from distribution of $\omega$
• $f(x) \approx (1/K) \sum_{i=1}^{K} f(x, \omega_i)$
• for each $i$ choose $g_i \in \partial_x f(x, \omega_i)$
• $g = (1/K) \sum_{i=1}^{K} g_i$ is an (approximate) subgradient (more on this later)
Minimization

define \( g(y) \) as the optimal value of

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq y_i, \quad i = 1, \ldots, m
\end{align*}
\]

\( (f_i \text{ convex}; \text{variable } x) \)

with \( \lambda^* \) an optimal dual variable, we have

\[
g(z) \geq g(y) - \sum_{i=1}^{m} \lambda_i^*(z_i - y_i)
\]

\( i.e., -\lambda^* \text{ is a subgradient of } g \text{ at } y \)
Composition

- \( f(x) = h(f_1(x), \ldots, f_k(x)) \), with \( h \) convex nondecreasing, \( f_i \) convex
- find \( q \in \partial h(f_1(x), \ldots, f_k(x)), g_i \in \partial f_i(x) \)
- then, \( g = q_1 g_1 + \cdots + q_k g_k \in \partial f(x) \)
- reduces to standard formula for differentiable \( h, f_i \)

proof:

\[
\begin{align*}
  f(y) &= h(f_1(y), \ldots, f_k(y)) \\
  \geq h(f_1(x) + g_1^T(y - x), \ldots, f_k(x) + g_k^T(y - x)) \\
  \geq h(f_1(x), \ldots, f_k(x)) + q^T(g_1^T(y - x), \ldots, g_k^T(y - x)) \\
  &= f(x) + g^T(y - x)
\end{align*}
\]
Subgradients and sublevel sets

$g$ is a subgradient at $x$ means $f(y) \geq f(x) + g^T(y - x)$

hence $f(y) \leq f(x) \implies g^T(y - x) \leq 0$
• $f$ differentiable at $x_0$: $\nabla f(x_0)$ is normal to the sublevel set
\[
\{ x \mid f(x) \leq f(x_0) \}
\]

• $f$ nondifferentiable at $x_0$: subgradient defines a supporting hyperplane
to sublevel set through $x_0$
Quasigradients

$g \neq 0$ is a **quasigradient** of $f$ at $x$ if

$$g^T(y - x) \geq 0 \implies f(y) \geq f(x)$$

holds for all $y$

quasigradients at $x$ form a cone
example:

\[ f(x) = \frac{a^T x + b}{c^T x + d}, \quad (\text{dom } f = \{x \mid c^T x + d > 0\}) \]

\[ g = a - f(x_0)c \] is a quasigradient at \( x_0 \)

proof: for \( c^T x + d > 0 \):

\[ a^T (x - x_0) \geq f(x_0) c^T (x - x_0) \implies f(x) \geq f(x_0) \]
**example:** degree of $a_1 + a_2 t + \cdots + a_n t^{n-1}$

$$f(a) = \min \{ i \mid a_{i+2} = \cdots = a_n = 0 \}$$

$g = \text{sign}(a_{k+1}) e_{k+1}$ (with $k = f(a)$) is a quasigradient at $a \neq 0$

**proof:**

$$g^T (b - a) = \text{sign}(a_{k+1}) b_{k+1} - |a_{k+1}| \geq 0$$

implies $b_{k+1} \neq 0$
Optimality conditions — unconstrained

recall for $f$ convex, differentiable,

$$f(x^*) = \inf_x f(x) \iff 0 = \nabla f(x^*)$$

generalization to nondifferentiable convex $f$:

$$f(x^*) = \inf_x f(x) \iff 0 \in \partial f(x^*)$$
proof. by definition (!)

\[ f(y) \geq f(x^*) + 0^T(y - x^*) \text{ for all } y \iff 0 \in \partial f(x^*) \]

\ldots seems trivial but isn't
Example: piecewise linear minimization

\[
f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i)
\]

\(x^\ast\) minimizes \(f \iff 0 \in \partial f(x^\ast) = \text{Co}\{a_i \mid a_i^T x^\ast + b_i = f(x^\ast)\}\)

\(\iff\) there is a \(\lambda\) with

\[
\lambda \succeq 0, \quad 1^T \lambda = 1, \quad \sum_{i=1}^{m} \lambda_i a_i = 0
\]

where \(\lambda_i = 0\) if \(a_i^T x^\ast + b_i < f(x^\ast)\)
... but these are the KKT conditions for the epigraph form

\[
\begin{align*}
\text{minimize} & \quad t \\
\text{subject to} & \quad a_i^T x + b_i \leq t, \quad i = 1, \ldots, m
\end{align*}
\]

with dual

\[
\begin{align*}
\text{maximize} & \quad b^T \lambda \\
\text{subject to} & \quad \lambda \succeq 0, \quad A^T \lambda = 0, \quad 1^T \lambda = 1
\end{align*}
\]
Optimality conditions — constrained

minimize $f_0(x)$
subject to $f_i(x) \leq 0, \ i = 1, \ldots, m$

we assume

• $f_i$ convex, defined on $\mathbb{R}^n$ (hence subdifferentiable)
• strict feasibility (Slater’s condition)

$x^*$ is primal optimal ($\lambda^*$ is dual optimal) iff

$$
\begin{align*}
f_i(x^*) &\leq 0, \quad \lambda_i^* \geq 0 \\
0 &\in \partial f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \partial f_i(x^*) \\
\lambda_i^* f_i(x^*) & = 0
\end{align*}
$$

... generalizes KKT for nondifferentiable $f_i$
Directional derivative

directional derivative of $f$ at $x$ in the direction $\delta x$ is

$$f'(x; \delta x) \triangleq \lim_{h \to 0} \frac{f(x + h\delta x) - f(x)}{h}$$

can be $+\infty$ or $-\infty$

- $f$ convex, finite near $x \implies f'(x; \delta x)$ exists

- $f$ differentiable at $x$ if and only if, for some $g (= \nabla f(x))$ and all $\delta x$, $f'(x; \delta x) = g^T \delta x$ (i.e., $f'(x; \delta x)$ is a linear function of $\delta x$)
Directional derivative and subdifferential

general formula for convex $f$: $f'(x; \delta x) = \sup_{g \in \partial f(x)} g^T \delta x$
Descent directions

$\delta x$ is a **descent direction** for $f$ at $x$ if $f'(x; \delta x) < 0$

for differentiable $f$, $\delta x = -\nabla f(x)$ is always a descent direction (except when it is zero)

**warning:** for nondifferentiable (convex) functions, $\delta x = -g$, with $g \in \partial f(x)$, need not be descent direction

eexample: $f(x) = |x_1| + 2|x_2|$
Subgradients and distance to sublevel sets

if $f$ is convex, $f(z) < f(x)$, $g \in \partial f(x)$, then for small $t > 0$,

$$\|x - tg - z\|_2 < \|x - z\|_2$$

thus $-g$ is descent direction for $\|x - z\|_2$, for any $z$ with $f(z) < f(x)$ (e.g., $x^*$)

negative subgradient is descent direction for distance to optimal point

proof: $\|x - tg - z\|^2_2 = \|x - z\|^2_2 - 2tg^T(x - z) + t^2\|g\|^2_2$

$\leq \|x - z\|^2_2 - 2t(f(x) - f(z)) + t^2\|g\|^2_2$
Descent directions and optimality

**fact:** for \( f \) convex, finite near \( x \), either

- \( 0 \in \partial f(x) \) (in which case \( x \) minimizes \( f \)), or

- there is a descent direction for \( f \) at \( x \)

*i.e.*, \( x \) is optimal (minimizes \( f \)) iff there is no descent direction for \( f \) at \( x \)

**proof:** define \( \delta x_{sd} = -\arg\min_{z \in \partial f(x)} \|z\|_2 \)

if \( \delta x_{sd} = 0 \), then \( 0 \in \partial f(x) \), so \( x \) is optimal; otherwise

\[ f'(x; \delta x_{sd}) = -\left(\inf_{z \in \partial f(x)} \|z\|_2\right)^2 < 0, \] so \( \delta x_{sd} \) is a descent direction
idea extends to constrained case (feasible descent direction)