Optimization for machine learning

Rokhlin D.B.

IMMCS, SFU, 2021

General optimization problem, local and global minimum

$$f(x) \to \min_{x \in S}, \quad S \subset \mathbb{R}^n.$$

- $x^* \in S$ is called a *global minimum* of (f on S), if $f(x^*) \le f(x)$, $x \in S$.

$$\|x-y\|^2 := \langle x-y, x-y \rangle := \sum_{i=1}^{n} (x^i - y^i)^2,$$

$$B_\varepsilon(x) = \{y \in \mathbb{R}^n : \|y - x\| < \varepsilon\}$$

is an open ball in $\mathbb{R}^n$.

- $x^* \in S$ is called a *local minimum* (f on S), if there exists $\varepsilon > 0$ such that

$$f(x^*) \le f(x), \quad x \in S \cap B_\varepsilon(x^*),$$

- $x^* \in S$ is called a *strict local minimum*, if there exists $\varepsilon > 0$ such that

$$f(x^*) < f(x), \quad x \in S \cap \mathring{B}_\varepsilon(x^*), \quad \mathring{B}_\varepsilon(x^*) := B_\varepsilon(x^*) \backslash \{x^*\}.$$

Taylor formula

Let $f \in C^2(\mathbb{R}^n)$, then

$$\begin{aligned} f(x) &= f(y) + \langle f'(y), x-y \rangle + o(\|x-y\|) \\ &= f(y) + \langle f'(y), x-y \rangle + \frac{1}{2}\langle f''(y)(x-y), x-y \rangle + o(\|x-y\|^2). \end{aligned}$$

Here

- $f'(y) = (x_{x^j}(y))_{j=1}^n$ is the gradient,
- $f''(y) = (x_{x^i x^j}(y))_{i,j=1}^n$ is the Hessian,
- $g(y) = o(\|y-x\|^\alpha)$, $y \to x$, if $g(y)/\|y-x\|^\alpha \to 0$, $y \to x$.

Optimality conditions in the unconstrained optimization problem'

$$f(x) \to \min_{x \in \mathbb{R}^n}.$$

Theorem (Necessary optimality conditions)

Let x^* be a local minimum, then

$$f'(x^*) = 0; \quad \langle f''(x^*)h, h \rangle \geq 0, \quad h \in \mathbb{R}^n.$$

Theorem (Sufficient optimality conditions)

Let

$$f'(x^*) = 0; \quad \langle f''(x^*)h, h \rangle > 0, \quad 0 \neq h \in \mathbb{R}^n,$$

then x^* is a strict local minimum.

Sylvester's criterion

Square symmetric $n \times n$ matrix matrix G is called *positive definite*, if

$$\langle Gh, h \rangle > 0, \quad 0 \neq h \in \mathbb{R}^n,$$

positive semidefinite, if

$$\langle Gh, h \rangle \geq 0, \quad h \in \mathbb{R}^n.$$

Denote by $\Delta_{i_1,\ldots,i_k}$ the *principal minor* of G, corresponding to a submatrix with identical numbers $i_1, \ldots, i_k$ of rows and columns. *Leading principal minors* are of the form $\Delta_{1,\ldots,k}$.

Theorem (Sylvester's criterion)

- *G is positive definite $\iff$ all leading principal minors are positive.*
- *G is positive semidefinite $\iff$ all principal minors are non-negative.*

Example: least squares method

- Find the global minimum point

$$\|Ax - b\|^2 \to \min_{x \in \mathbb{R}^n},$$

assuming that the columns of A are linearly independent.

Consider the function

$$f(x) = \|Ax - b\|^2 = \langle Ax - b, Ax - b \rangle = \langle A^T A x, x \rangle - 2\langle A^T b, x \rangle + \|b\|^2,$$

and find its gradient and Hessian:

$$f'(x) = 2A^T A x - 2A^T b, \quad f''(x) = 2A^T A.$$

If x^* is a local minimum point, then

$$A^T A x^* = A^T b.$$

Square matrix $A^T A$ is invertible, since if $A^T Ay = 0$, then $\langle A^T Ay, y\rangle = \|Ay\| = 0$ and $y = 0$. Indeed, if

$$Ay = \sum_{j=1}^{n} A_j y_j = 0,$$

then $y = 0$ by the linear independence of the columns of A. Thus there is at most one local minimum point

$$x^* = (A^T A)^{-1} A^T b.$$

Hessian $f''(x^*)$ is positive definite:

$$\frac{1}{2}\langle f''(x^*)h, h\rangle = \langle A^T Ah, h\rangle = \|Ah\|^2 \geq 0,$$

and the equality $Ah = 0$ implies that $h = 0$. Hence, x^* is a strict local minimum.

Moreover, it is a global minimum, since

$$\begin{aligned} f(x^* + y) &= \langle A^T A(x^* + y), x^* + y\rangle - 2\langle A^T b, x^* + y\rangle + \|b\|^2 \\ &= f(x^*) + \langle A^T Ay, y\rangle + \langle A^T Ax^*, y\rangle + \langle A^T Ay, x^*\rangle - 2\langle A^T b, y\rangle \\ &= f(x^*) + \langle A^T Ay, y\rangle > f(x^*), \quad y \neq 0. \end{aligned}$$

In fact, this assertion also follows from the convexity of f.

Optimization under constraints

Theorem (necessary Karush-Kuhn-Tucker conditions)

Let $\overline{x}$ be a local minimum point in the problem

$$f(x) \to \min, \quad S : g_i(x) \leq 0,\ i = 1, \ldots, m, \quad h_i(x) = 0,\ i = 1, \ldots, k.$$

Put

$$L(x, \lambda, mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{k} \mu_i h_i(x).$$

Under some technical regularity conditions there exist $\overline{\lambda}_i \geq 0$, $i = 1, \ldots, m$, $\overline{\mu}_i$, $i = 1, \ldots, k$ *such that the stationarity and complementary slackness conditions are satisfied:*

$$L_x(\overline{x}, \overline{\lambda}, \overline{\mu}) = 0; \quad \overline{\lambda}_i g_i(\overline{x}) = 0,\ i = 1, \ldots, m.$$

Convex functions

A set $G \subset \mathbb{R}^d$ is called convex if $\alpha x + (1-\alpha)y \in G$ for all $x, y \in G$, $\alpha \in [0,1]$. A function $f : G \mapsto \mathbb{R}$, defined on a convex set G, is called

- convex if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad x, y \in G, \alpha \in [0,1]$$

- λ-strongly convex ($\lambda \geq 0$) if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\lambda}{2}\alpha(1-\alpha)\|x-y\|^2$$

  for all $\alpha \in [0,1]$, $x, y \in \mathbb{R}^n$.

A convex function satisfies the Jensen inequality:

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i f(x_i), \quad \alpha_i \geq 0, \quad \sum_{i=1}^n \alpha_i = 1.$$

Convexity criteria

A function, defined on a convex set G is convex if and only if its epigraph

$$\operatorname{epi} f = \{(x, y) \in G \times \mathbb{R} : f(x) \leq y\}$$

is convex.

Theorem

- *Assume that f is differentiable in a neighborhood of a convex set G. Then $f : G \mapsto \mathbb{R}$ is convex if and only if*

$$f(x) \geq f(y) + \langle f'(y), x - y \rangle, \quad x, y \in G.$$

- *Assume that a convex set G is open. Twice continuously differentiable function $f : G \mapsto \mathbb{R}$ is convex if and only if*

$$\langle f''(x)h, h \rangle \geq 0, \quad h \in \mathbb{R}^d, \quad x \in G.$$

Examples

- $f(x) = \|Ax - b\|^2$ is convex since its Hessian $f''(x) = 2A^T A$ is positive semidefinite.
- $f(x_1, x_2) = x_1^2/x_2$ is convex on a set $\{x : x_2 > 0\}$, since the principal minors of the Hessian

$$f''(x) = \begin{pmatrix} 2/x_2 & -2x_1/x_2^2 \\ -2x_1/x_2^2 & 2x_1^2/x_2^3 \end{pmatrix} = \frac{2}{x_2^2} \begin{pmatrix} x_2 & -x_1 \\ -x_1 & x_1^2/x_2 \end{pmatrix}$$

are non-negative: $x_2 > 0$, $x_1^2/x_2 > 0$, $x_2 x_1^2/x_2 - x_1^2 = 0$.

Strong convexity criteria

Theorem

Let G be an open set.

- *A differentiable function $f : G\mathbb{R}^d \mapsto \mathbb{R}$ is μ-strongly convex if and only if*
$$f(x) \geq f(y) + \langle f'(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2.$$
- *A twice differentiable function $f : G \mapsto \mathbb{R}$ is μ-strongly convex, if and only if*
$$\langle f''(x)h, h\rangle \geq \mu\|h\|^2 I, \; x \in G.$$

Assume that the columns of A a linearly independent. Then $A^T A$ is positive definite. Let $\lambda_1 > 0$ be its smallest eigenvalue. Then $f(x) = \|Ax - b\|^2$ is $2\lambda_1$-strongly convex:

$$\langle f''(x)h, h\rangle = 2\langle A^T Ah, h\rangle \geq 2\lambda_1\|h\|^2.$$

Operations, preserving convexity

(a) *Conical combination.* Let f_i be convex and $\lambda_i \geq 0$, then the function $f(x) = \lambda_1 f_1(x) + \cdots + \lambda_n f_n(x)$ is convex.

(b) *Affine substitution.* Let f be convex, then $g(x) = f(Ax+b)$ is convex.

(c) *Maximization over a parameter.* Let $x \mapsto f(x, y)$ be convex for any $y \in Y$. Then $x \mapsto g(x) = \sup_{y \in Y} f(x, y)$ is convex. Indeed,

$$\operatorname{epi} g = \cap_{y \in Y} \operatorname{epi} f(\cdot, y).$$

(d) *Minimization over a parameter.* Let f be convex in (x, y), and S be a non-empty convex set. Then the function

$$g(x) = \inf_{y \in S} f(x, y)$$

is convex, if $g(x) > -\infty$ for some x.

(e) *Superposition.* Let

$$f(x) = h(g(x)) = h(g_1(x), \dots, g_k(x)),$$

where $h : \mathbb{R}^k \mapsto \mathbb{R}$ is convex. Then f is convex under any of the following conditions:

(i) h is convex in each argument and the functions $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ are convex;

(ii) h is non-increasing in each argument and the functions $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ are concave.

To understand this property consider the one-dimensional case $k = d = 1$ and assume that the functions h and g are twice differentiable. Then

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x).$$

The first term is non-negative by the convexity of f. The second term is non-negative, since in both cases (i), (ii) $h'(g(x))$ and $g''(x)$ have the same sign.

Example

The support function of a set S

$$\sigma_S(x) = \sup_{y \in S} \langle x, y \rangle,$$

is convex as a pointwise maximum of a family of linear functions $x \mapsto \langle x, y \rangle$.

Example

The distance from a point x to a convex set S

$$f(x) = \inf_{y \in S} \|x - y\|,$$

is convex, since $\|x - y\|$ is convex in (x, y) as a superposition of a convex function $\| \cdot \|$ and an affine function.

Any local minimum of a convex function is global

$x^* \in \mathbb{G}$ is called a *global minimum* of $f : G \mapsto \mathbb{R}$, if $f(x^*) \leq f(x)$, $x \in G$. $x^* \in \mathbb{R}^d$ is called a *local minimum* of $f : G \mapsto \mathbb{R}$, if there exists $\varepsilon > 0$ such that

$$f(x^*) \leq f(x), \quad x \in B_\varepsilon(x^*) := \{y : \|y - x\| \leq \varepsilon\}.$$

Theorem

Any local minimum point $x^* \in \operatorname{dom} f$ of a convex function f is its global minimum.

Proof.

For any point $x \in \mathbb{R}^d$ select a sufficiently small $\alpha \in (0, 1)$ such that $f(x^*) \leq f(x^* + \alpha(x - x^*))$. Using the convexity of f, we get

$$f(x^*) \leq f(x^* + \alpha(x - x^*)) = f(\alpha x + (1 - \alpha)x^*) \leq \alpha f(x) + (1 - \alpha)f(x^*).$$

It follows that $f(x^*) \leq f(x)$. $\square$

Optimality condition

Theorem

Let $f : G \mapsto \mathbb{R}$ be a convex function. $w^* \in G$ is a (global) minimum of f if and only if

$$\langle f'(w^*), w - w^* \rangle \geq 0, \quad w \in G.$$

Proof. If w^* is a local minimum, then for any $w \in G$ we have

$$0 \leq f(w^* + \alpha(w - w^*)) - f(w^*) = \alpha \langle f'(w^*), w - w^* \rangle + o(\alpha)$$

for sufficiently small $\alpha > 0$. Here we used the convexity of G: $w^* + \alpha(w - w^*) \in G$. It follows that

$$\langle f'(w^*), w - w^* \rangle \geq 0, \quad w \in G.$$

Conversely, if the last inequality holds true, then w^* is the global minimum:

$$f(w) - f(w^*) \geq \langle f'(w^*), w - w^* \rangle \geq 0, \quad w \in G. \quad \square$$

Projection theorem

A point $\Pi_G(w) = \arg\min_{u \in G} \|w - u\|$ is called the projection of w onto G:

$$\|\Pi_G(w) - w\| \leq \|u - w\|, \quad u \in G.$$

In general a projection need not exist or be unique.

Theorem

If G is closed and convex, then there exists a unique projection $\Pi_G(w)$. It is characterized by the inequality:

$$\langle u - \Pi_G(w), w - \Pi_G(w) \rangle \leq 0, \quad u \in G.$$

Proof. By the definition, $\Pi_G(w)$ is the minimum point of

$$f(u) = \|u - w\|^2/2$$

on G. The existence can be deduced from an appropriately modified Weierstrass theorem: the set G need not be bounded but $f(u) \to \infty$, $\|u\| \to \infty$. The uniqueness follows from the strong convexity of f:

$$f'(u) = u - w, \quad f''(u) = I.$$

The optimality condition for $\Pi_G(w)$ gives the inequality

$$\langle f'(\Pi_G(w)), u - \Pi_G(w)\rangle = \langle \Pi_G(w) - w, u - \Pi_G(w)\rangle \geq 0. \quad \square$$

The projection is non-expansive

Lemma

The projection operator in non-expansive:

$$\|\Pi_G(w) - \Pi_G(u)\| \leq \|w - u\|.$$

Proof. By the projection theorem,

$$\langle \Pi_G u - \Pi_G w, w - \Pi_G w \rangle \leq 0,$$

$$\langle \Pi_G w - \Pi_G u, u - \Pi_G u \rangle \leq 0,$$

Summing up, we get

$$\langle w - u - (\Pi_G w - \Pi_G u), \Pi_G u - \Pi_G w \rangle \leq 0,$$

$$\|\Pi_G u - \Pi_G w\|^2 \leq \langle w - u, \Pi_G w - \Pi_G u \rangle \leq \|w - u\| \|\Pi_G w - \Pi_G u\|.$$

Thus, $\|\Pi_G w - \Pi_G u\| \leq \|w - u\|$. $\square$

Kuhn-Tucker theorem

$$f(x) \to \min$$

$$S : g_i(x) \le 0, \quad i = 1, \dots, m.$$

Lagrange function:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x).$$

Assume that f, g_i are convex and the Slater condition is satisfied:

$$\exists \overline{x} \in S : g_i(\overline{x}) < 0, \quad i = 1, \dots, m.$$

A point $x^* \in S$ is an optimal solution (global minimum point) if and only if there exist $\lambda_i^* \ge 0$, $i = 1, \dots, m$ such that

$$L_{x_j}(x^*, \lambda^*) = 0, \quad j = 1, \dots, d, \quad \text{(stationarity)},$$

$$\lambda_i^* g_i(x^*) = 0, \quad i = 1, \dots, m, \quad \text{(complementary slackness)}.$$

Subdifferentials

A function $f : G \mapsto \mathbb{R}$ is called *subdifferentiable* at w, if there exists $\gamma \in \mathbb{R}^n$ such that

$$f(u) \geq f(w) + \langle \gamma, u - w \rangle, \quad u \in G.$$

γ is called a *subgradient* of f at w. The set $\partial f(w)$ of all subgradients is called a *subdifferential*.

- The set $\partial f(w)$ is closed and convex.
- The set $\partial f(w)$ can be empty.
- If f is differentiable and subdifferentiable, then $\partial f(w) = \{f'(w)\}$.

Lemma

If $\partial f(w) \neq \emptyset$, $w \in G$, then f is convex.

Proof. For $u, v \in G$ put $w = \alpha u + (1-\alpha)v$, $\alpha \in [0,1]$.

$$f(u) \geq f(w) + \langle \gamma, u - w \rangle = f(w) + (1-\alpha)\langle \gamma, u - v \rangle$$
$$f(v) \geq f(w) + \langle \gamma, v - w \rangle = f(w) + \alpha\langle \gamma, v - u \rangle$$

$$\alpha f(u) + (1-\alpha) f(v) \geq f(w). \quad \square$$

Theorem

Let $f : G \mapsto \mathbb{R}$ *be a convex function and* $x \in \operatorname{int}(\operatorname{dom} f)$. *Then*

- $\partial f(x)$ *is a non-empty convex set.*
- *If* f *is differentiable at* x, *then* $\partial f(x) = \{f'(x)\}$. *Conversely, if a subgradient at* x *is unique, then* f *is differentiable at* x, *and* $\partial f(x) = \{f'(x)\}$.

Theorem

Let $f_1, \ldots, f_m$ *be convex function defined on* $G_1, \ldots, G_n$. *If* $x \in G = \cap_{i=1}^n G_i$, *then*

$$\sum_{i=1}^m \partial f_i(x) \subseteq \partial \left(\sum_{i=1}^m f_i \right)(x).$$

If at some point $\overline{x} \in G$, *all function* $f_1, \ldots, f_m$, *except maybe one, are continuous, then*

$$\partial \left(\sum_{i=1}^m f_i \right)(x) = \sum_{i=1}^m \partial f_i(x), \quad x \in \mathbb{R}^d.$$

$$\text{conv } A = \left\{ \sum_{i=1}^{m} \alpha_i x_i : x_i \in A, \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i = 1, m \in \mathbb{N} \right\}.$$

Theorem

Consider a family of convex functions f_i, $i \in I$, defined on G_i, where I is an arbitrary index set. Let $f(x) = \sup_{i \in I} f_i(x)$. Then

$$\text{conv } \left(\cup_{i \in I(x)} \partial f_i(x) \right) \subseteq \partial f(x), \quad x \in G = \cap_{i \in I} G_i, \tag{1}$$

where $I(x) = \{i \in I : f_i(x) = f(x)\}$. If I is finite and at some point $\overline{x} \in G$ all functions are continuous, then

$$\partial f(x) = \text{conv } \left(\cup_{i \in I(x)} \partial f_i(x) \right), \quad x \in \mathbb{R}^d. \tag{2}$$

In particular, if $f(w) = \max_{i \in J} f_i(w)$, where J is a finite set and f_i are differentiable, and $j \in \arg\max_{i \in J} g_i(w)$, then $f'_j(w) \in \partial f(w)$.
Note that (1) is implied by the inequality

$$f(y) \geq f_i(y) \geq f_i(x) + \langle g, y - x \rangle = f(x) + \langle g, y - x \rangle, \quad g \in \partial f_i(x),$$

where $i \in I(x)$.

Theorem

Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ *be a convex function, and* $g : \mathbb{R} \mapsto \mathbb{R}$ *be non-decreasing convex function. Then* $h = g \circ f$ *is convex, and if* g *is differentable at* $f(x)$*, then*

$$\partial h(x) = g'(f(x))\partial f(x).$$

Theorem

If $f : G \mapsto \mathbb{R}$ *is a convex function,* $A : \mathbb{R}^k \mapsto \mathbb{R}^d$ *is a linear operator, and* $b \in \mathbb{R}^d$*. Put* $h(x) = f(Ax + b)$*. If*

$$H = \{x \in \mathbb{R}^k : Ax + b \in \operatorname{dom} f\} \neq \emptyset,$$

then

$$A^T \partial f(Ax + b) \subseteq \partial h(x), \quad x \in H,$$

and if f *is continuous at some point* $\overline{y} = A\overline{x} + b$*, then*

$$A^T \partial f(Ax + b) = \partial h(x), \quad x \in H. \tag{3}$$

For differentiable functions formula (3) is a consequence of Taylor's formula:

$$
\begin{aligned}
& f(A(x+\varepsilon\gamma)+b) - f(Ax+b) = f(Ax+b+\varepsilon A\gamma) - f(Ax+b) \\
& = \langle f'(Ax+b), \varepsilon A\gamma\rangle + o(\varepsilon) = \varepsilon\langle A^T f'(Ax+b), \gamma\rangle + o(\varepsilon), \quad \gamma \in \mathbb{R}^k.
\end{aligned}
$$

Example

Let $f(x) = |x| = \max\{-x, x\}$, $x \in \mathbb{R}$. Then $\partial f(0) = \operatorname{conv}\{-1, 1\} = [-1, 1]$. Thus,

$$
\partial f(x) = \begin{cases} 1, & x > 0 \\ [-1, 1], & x = 0, \\ -1, & x < 0. \end{cases}
$$

Example

$f(w) = \max\{0, 1 - y\langle w, x\rangle\}$,

$$v = \begin{cases} 0, & 1 - \langle w, x\rangle \leq 0 \\ -yx, & 1 - \langle w, x\rangle > 0 \end{cases}$$

belongs to $\partial f(w)$.

Example

Let $f(x) = \|x\|_1 = \sum_{i=1}^{d} f_i(x_i)$, $f_i(x_i) = |x_i|$. Then

$$\partial f_i(x_i) = \begin{cases} \{\operatorname{sgn}(x_i) e_i\}, & x_i \neq 0, \\ [-e_i, e_i], & x_i = 0, \end{cases}$$

where e_i is a vector of the standard basis of $\mathbb{R}^d$, that is i-th component of e_i equals to 1, and others equals to 0,

$$\operatorname{sgn}(y) = \begin{cases} 1, & y \geq 0, \\ -1, & y < 0. \end{cases}$$

By the formula for the subdifferential of a sum,

$$\partial f(x) = \sum_{i=1}^{d} \partial f_i(x_i) = \sum_{i \in I_1(x)} \operatorname{sgn}(x_i) e_i + \sum_{i \in I_0(x)} [-e_i, e_i],$$

$I_0(x) = \{i : x_i = 0\}$, $I_1(x) = \{i : x_i \neq 0\}$. in other words,

$$\partial f(x) = \{z \in \mathbb{R}^d : z_i = \operatorname{sgn}(x_i),\ i \in I_1(x);\ |z_i| \leq 1,\ i \in I_0(x)\}.$$

Example

Let $f(x) = \max_{1 \leq i \leq m}(\langle a_i, x\rangle + b_i)$. Then

$$\partial f(x) = \left\{ \sum_{i \in I(x)} \lambda_i a_i : \sum_{i \in I(x)} \lambda_i = 1, \lambda \geq 0 \right\},$$

where $I(x) = \{i : f(x) = \langle a_i, x\rangle + b_i\}$.

Example

Let $h(x) = \|Ax - b\|_1$, where $A \in \mathbb{R}^{d\times k}$, $b \in \mathbb{R}^d$. Introduce the index set

$$I_0(x) = \{i : \langle a_i, x\rangle = b_i\}, \quad I_1(x) = \{i : \langle a_i, x\rangle \neq b_i\},$$

where a_i are the rows of A. According to one of the previous examples, for $g(y) = \|y\|_1$ we have

$$\partial g(Ax + b) = \sum_{i\in I_1(x)} \operatorname{sgn}(\langle a_i, x\rangle - b_i)e_i + \sum_{i\in I_0(x)} [-e_i, e_i].$$

Hence, according to (3),

$$\partial h(x) = A^T \partial g(Ax + b) = \sum_{i\in I_1(x)} \operatorname{sgn}(\langle a_i, x\rangle - b_i)A^T e_i + \sum_{i\in I_0(x)} [-A^T e_i, A^T e_i]$$

$$= \sum_{i\in I_1(x)} \operatorname{sgn}(\langle a_i, x\rangle - b_i)a_i^T + \sum_{i\in I_0(x)} [-a_i^T, a_i^T].$$

Example

$$f(w) = |y - \langle w, x\rangle|_\varepsilon = \max\{0, |y - \langle w, x\rangle| - \varepsilon\}$$
$$= \max\{0, y - \langle w, x\rangle - \varepsilon, \langle w, x\rangle - y - \varepsilon\}$$

$$v = \begin{cases} 0, & |y - \langle w, x\rangle| \le \varepsilon \\ -x, & y \ge \langle w, x\rangle - \varepsilon \\ x, & y \le \langle w, x\rangle - \varepsilon \end{cases}$$

belongs to $\partial f(w)$.

Lemma (subdifferential of a Lipschitz function)

Let $f : G \mapsto \mathbb{R}$ be a convex function. if $\|v\| \le \rho$, $v \in \partial f(w)$, $w \in G$, then f is ρ-Lipschitz:

$$|f(u) - f(w)| \le \rho \|u - w\|.$$

If G is open then converse statement is also true.

Proof. " $\Longrightarrow$ " If $\|\gamma\| \le \rho$ for $\gamma \in \partial f(w)$ then

$$f(u) - f(w) \ge \langle \gamma, u - w \rangle,$$

$$f(w) - f(u) \le \langle \gamma, w - u \rangle \le \|\gamma\| \|w - u\| \le \rho \|w - u\|.$$

Similarly,

$$f(u) - f(w) \le \rho \|w - u\|.$$

Hence, f is ρ-Lipschitz.
" $\Longleftarrow$ " Assume than f is ρ-Lipschitz. We want to prove that $\|\gamma\| \le \rho$, $\gamma \in \partial f(w)$, $w \in G$. For sufficiently small $\varepsilon > 0$ we have

$$u = w + \varepsilon \frac{\gamma}{\|\gamma\|} \in G,$$

$$f(u) - f(w) \ge \langle \gamma, u - w \rangle = \langle \gamma, \varepsilon \frac{\gamma}{\|\gamma\|} \rangle = \varepsilon \|\gamma\|.$$

But

$$f(u) - f(w) \le \rho \|u - w\| = \rho \varepsilon$$

since f is ρ-Lipschitz. Thus, $\|\gamma\| \le \rho$. $\square$

Projection, (projected) subgradient descent method

Consider the convex minimization problem

$$f(w) \to \min_{w \in G}.$$

(Projected) (sub)gradient descent (GD) method:

$$w_{t+1} = \Pi_G(w_t - \eta_t v_t), \quad v_t \in \partial f(w_t).$$

A basic inequality

Put $r_t = \|w_t - w^*\|$, where w^* is a global minimum point of f over G. We have

$$\begin{aligned} r_{t+1}^2 - r_t^2 &= \|\Pi_G(w_t - \eta_t v_t) - \Pi_G w^*\|^2 - \|w_t - w^*\|^2 \\ &\le \|w_t - w^* - \eta_t v_t\|^2 - \|w_t - w^*\|^2 \\ &= -2\eta_t \langle w_t - w^*, v_t\rangle + \eta_t^2 \|v_t\|^2. \end{aligned}$$

By the definition of a subgradient,

$$f(w^*) - f(w_t) \ge \langle v_t, w^* - w_t\rangle = -\langle v_t, w_t - w^*\rangle.$$

Thus,

$$r_{t+1}^2 - r_t^2 \le -2\eta_t (f(w_t) - f(w^*)) + \eta_t^2 \|v_t\|^2.$$

Theorem

Let $f : G \mapsto \mathbb{R}$ be a convex ρ-Lipschitz function. Then

$$f_T^* - f^* \le \frac{1}{2}\frac{r_1^2 + \rho^2 \sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t}, \tag{4}$$

where $f_T^* = \min_{1 \le t \le T} f(w_t)$, $f^* = f(w^*)$, $r_1 = \|w_1 - w^*\|$.

Proof. Summing up the basic inequalities over $t = 1, \dots, T$, we get

$$\sum_{t=1}^T 2\eta_t(f(w_t) - f^*) \le \sum_{t=1}^T (r_t^2 - r_{t+1}^2) + \sum_{t=1}^T \eta_t^2 \|u_t\|^2$$

$$\le r_1^2 - r_{T+1}^2 + \rho^2 \sum_{t=1}^T \eta_t^2 \le r_1^2 + \rho^2 \sum_{t=1}^T \eta_t^2 .$$

The result follows in an evident way. □

- The estimate in the last theorem holds true not only for the best point w_T^*: $f(w_T^*) = f_T^*$, but also for the average approximation

$$\overline{w}_T = \sum_{t=1}^{T} \frac{\eta_t}{\sum_{j=1}^{T} \eta_j} w_t.$$

Indeed, by Jensen's inequality,

$$f(\overline{w}_T) - f^* \leq \sum_{t=1}^{T} \frac{\eta_t}{\sum_{j=1}^{T} \eta_j} (f(w_t) - f^*) \leq \frac{1}{2} \frac{B^2 + \rho^2 \sum_{t=1}^{T} \eta_t^2}{\sum_{t=1}^{T} \eta_t}$$

For a constant step size $\eta_t = \eta$, $\overline{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$.

- Instead of the Lipschitz condition it is enough to require that the inequality $\|g_t\|_2 \leq L$ is satisfied only for the subgradients $g_t \in \partial f(x_t)$, used in the algorithm.

An "optimal" constant step size

Assume that $r_1 \leq B$. For a constant step size $\eta_t = \eta$ the inequality in the theorem takes the form

$$f_T^* - f^* \leq \frac{1}{2}\frac{B^2}{T\eta} + \frac{\eta\rho^2}{2}.$$

The minimum of the right-hand side is attained at

$$\eta = \frac{B}{\rho\sqrt{T}}.$$

For this step we have

$$f_T^* - f^* \leq \frac{B\rho}{\sqrt{T}}.$$

Time-varying step size

The mentioned constant step size depends on the number T of iterations. We get almost the same estimate using the time-varying step size

$$\eta_t = \frac{B}{\rho\sqrt{t}},$$

which does not depend on T. This follows from the Theorem since

$$\sum_{t=1}^{T} t^{-1/2} \sim \sqrt{T}, \quad \sum_{t=1}^{T} 1/t \sim \ln T, \quad T \to \infty.$$

Let us get more precise estimates.

Lemma

Let $a \leq b$ be some integers. For a continuous non-increasing function $f : [a-1, b+1] \mapsto \mathbb{R}$ we have

$$\int_a^{b+1} f(x)\,dx \leq f(a) + f(a+1) + \cdots + f(b) \leq \int_{a-1}^{b} f(x)\,dx.$$

Proof.

$$\int_a^{b+1} f(x)\,dx = \int_a^{a+1} f(x)\,dx + \cdots + \int_b^{b+1} f(x)\,dx \leq f(a) + \cdots + f(b),$$

$$\int_{a-1}^{b} f(x)\,dx = \int_{a-1}^{a} f(x)\,dx + \cdots + \int_{b-1}^{b} f(x)\,dx \geq f(a) + \cdots + f(b). \qquad \square$$

It follows that

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \int_{1}^{T+1} \frac{dx}{\sqrt{x}} = 2(\sqrt{T+1} - 1), \tag{5}$$

$$\sum_{t=1}^{T} \frac{1}{t} = 1 + \sum_{t=2}^{T} \frac{1}{t} \leq 1 + \int_{1}^{T} \frac{dx}{x} = 1 + \ln T. \tag{6}$$

Thus,

$$\sum_{t=1}^{T} \eta_t \geq \frac{B}{\rho} \frac{\sqrt{T+1} - 1}{2}, \quad \sum_{t=1}^{T} \eta_t^2 \leq \frac{B^2}{\rho^2} (\ln T + 1)$$

and the Theorem implies

$$f_T^* - f^* \leq \rho \frac{B^2 + B^2(\ln T + 1)}{B(\sqrt{T+1} - 1)} = B\rho \frac{\ln T + 2}{\sqrt{T+1} - 1}.$$

A more subtle reasoning allows to get rid of $\ln T$ in the numerator.

Example

Let $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, n$ be a dataset. Consider l^1-linear regression problem with an additional constraint, imposed on the coefficients:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} |\langle x_i, w\rangle - y_i| = \frac{1}{n}\|Aw - b\|_1 \to \min_{\|w\|_p \le \rho}, \quad p = 2 \;\text{ or }\; p = \infty.$$

Here $A \in \mathbb{R}^{n\times d}$ is the matrix with rows $x_i \in \mathbb{R}^d$, and $b = (y_1, \dots, y_n) \in \mathbb{R}^n$. As was showed above,

$$g = \frac{1}{n}\sum_{i\in I_1(w)} \operatorname{sgn}\left(\langle x_i, w\rangle - y_i\right) x_i^T \in \partial f(w), \tag{7}$$

$I_1(w) = \{i : \langle x_i, w\rangle \neq y_i\}$. Hence, the subgradient descent method takes the form

$$w_{t+1} = \Pi_S\left(w_t - \frac{\eta_t}{n}\sum_{i\in I_1(w_t)} \operatorname{sgn}\left(\langle x_i, w_t\rangle - y_i\right) x_i^T\right),$$

where Π_S is the projection on the ball B_ρ of the space l^2:

$$\Pi_S(z) = \frac{z}{\max\{1, \|z\|_2/\rho\}} \tag{8}$$

or l^∞:

$$(\Pi_S(z))_i = \begin{cases} \rho, & z_i \geq 1, \\ z_i, & z_i \in (-1, 1), \\ -\rho, & z_i \leq -1, \end{cases} \qquad i = 1, \ldots, d. \tag{9}$$

The subgradients used in the algorithm satisfy the inequality

$$\|g\|_2 \leq L = \frac{1}{n} \sum_{i=1}^{n} \|x_i\|_2.$$