Ch. 4. Optimality, Convexity

We have seen that many data analysis tasks are formulated as an optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{or} \qquad \begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{s.t. } g_i(x)\le 0 \quad i=1,2,\cdots,p \\ &\qquad\quad h_i(x)=0 \quad i=1,2,\cdots,q. \end{aligned}$$

This chapter studies the optimality condition for these optimization problems. We consider optimization in $\mathbb{R}^n$ for simplicity. Most of the results in this chapter can be extended to Hilbert spaces V.

§ 4.1 Smooth Unconstrained Optimization.

Consider unconstrained optimization

$$\min_{x\in V} f(x). \qquad \text{(OPT)}$$

We assume $f(x)$ is differentiable.

- Solvability of (OPT).

  We say $x^{(*)}$ is a solution of (OPT) if

  $$f(x^{(*)}) \le f(x) \qquad \forall\, x\in V.$$

  In this case, we write $x^{(*)} = \arg\min_{x\in V} f(x)$.

  We also call $x^{(*)}$ a global minimizer of f in V.

  - The existence of a solution of (OPT) is NOT guaranteed.

    Example:

$f(x) = x^3, \quad x\in\mathbb{R}$

$f(x) = \langle a, x\rangle, \quad a, x\in V.$

$\vdots$

- When there exists a solution of (OPT)?

Thm: If f is continous (i.e., $f(x^{(n)}) \to f(x)$ as $x^{(n)} \to x$)
and coecive (i.e., $f(x^{(n)}) \to +\infty$ as $\|x^{(n)}\| \to +\infty$)

Then there exists at least a solution $x^{(*)}$ of (OPT).

proof. Consider the sets $S_\alpha = \{x \mid f(x) \le \alpha\}$ for a given $\alpha \in \mathbb{R}$.

S_α is closed: Let $x^{(n)} \to x$ and $\{x^{(n)}\} \subset S_\alpha$. Then $f(x^{(n)}) \le \alpha$.
By the continuity, $f(x) \le \alpha$. $\Rightarrow x \in S_\alpha$.

S_α is bounded: Suppose S_α is unbounded, i.e., $\exists \{x^{(n)}\} \subset S_\alpha$ s.t.
$\|x^{(n)}\| \to +\infty$. Then coecivity implies $f(x^{(n)}) \to +\infty$.
This contradicts with $f(x^{(n)}) \le \alpha$.

Therefore, S_α is bounded and closed for any $\alpha \in \mathbb{R}$.

We choose α s.t. S_α is non-empty. By Weierstrass's theorem, any continuous function on a bounded and closed set must have a minimizer, So, $\min_{x \in S_\alpha} f(x)$ has a solution, which is also a solution of $\min_{x \in \mathbb{R}^n} f(x)$. $\boxtimes$

- Necessary condition for optimality.

Theorem: Assume f is differentiable at $x^{(*)}$. Then,

$$x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x) \implies \nabla f(x^{(*)}) = 0$$

proof. By expansion,

$$f(x) = f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle + o(\|x - x^{(*)}\|)$$

Suppose $\nabla f(x^{(*)}) \neq 0$.

Then choose $\tilde{x} = x^{(*)} - t \nabla f(x^{(*)})$ with $t > 0$ gives

$$f(\tilde{x}) = f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + o(|t| \|\nabla f(x^{(*)})\|)$$

Since $t \|\nabla f(x^{(*)})\| \cdot \|\nabla f(x^{(*)})\|$, by choosing a sufficiently small t,

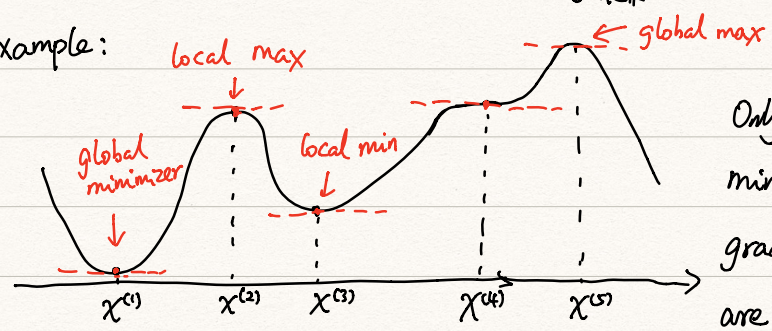
$$t\|\nabla f(x^{(*)})\|^2 > o\left(|t|\,\|\nabla f(x^{(*)})\|\right),$$

This implies $f(\tilde{x}) < f(x^{(*)})$, which contradicts with $x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$. $\boxtimes$

The condition $\nabla f(x^{(*)})=0$ is only a necessary condition.

The reverse " $\nabla f(x^{(*)})=0 \Rightarrow x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$ " is generally NOT true.

Example:

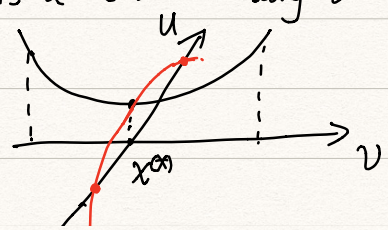


Only $x^{(1)}$ is a global minimizer, though the gradient at $x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}$ are also 0.

From this example, we see that $x^{(*)}$ with $\nabla f(x^{(*)})=0$ can be

- Global minimizer. i.e., $x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$ (see $x^{(1)}$)
- Local minimizer, i.e.,
  $\exists\, \varepsilon$, s.t. $f(x^{(*)}) \le f(x)$ $\forall\, x: \|x - x^{(*)}\| \le \varepsilon$. (see $x^{(2)}$)
- global max, i.e., $\forall\, x\in\mathbb{R}^n$, $f(x^{(*)}) \ge f(x)$, (See $x^{(5)}$)
- Local max, i.e, $\exists\, \varepsilon$. s.t. $f(x^{(*)}) \ge f(x)$ $\forall\, x: \|x - x^{(*)}\| \le \varepsilon$ (See $x^{(3)}$)
- Saddle points (only for $\mathbb{R}^n$ with $n \ge 2$), i.e.,
  $\exists\, u, v \in \mathbb{R}^n$ s.t. $f(x^{(*)}) \ge f(x^{(*)} + tu)$ for all $|t| \le \varepsilon$.
  and $f(x^{(*)}) \le f(x^{(*)} + tv)$ for all $|t| \le \varepsilon$.

(i.e., $f(x^{(*)})$ is a local min along v and a local max along u)

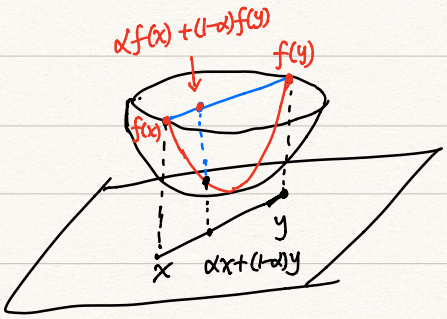


- None of the above (see $x^{(4)}$)

- Sufficient condition for optimality

Convexity : A function $f: \mathbb{R}^n \to \mathbb{R}$ is convex if

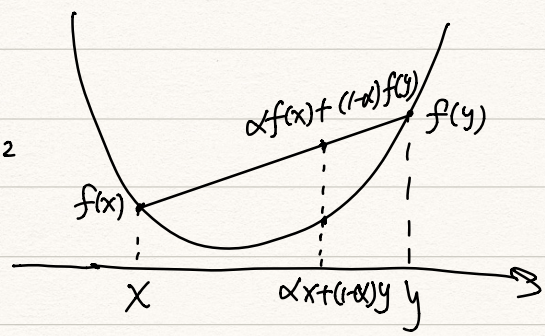
$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) \quad \forall\, x, y \in \mathbb{R}^n,\ \alpha \in [0,1].$$



The secant connecting $(x, f(x))$, $(y, f(y))$ is above the function graph.

Example 1: $f(x) = x^2$, $x \in \mathbb{R}$

$$\begin{aligned}
&f(\alpha x + (1-\alpha) y) \\
&= (\alpha x + (1-\alpha) y)^2 \\
&= \alpha^2 x^2 + 2\alpha(1-\alpha) x y + (1-\alpha)^2 y^2 \\
&= \alpha x^2 + (1-\alpha) y^2 + (\alpha^2 - \alpha) x^2 \\
&\quad + (\alpha^2 - \alpha) y^2 + 2\alpha(1-\alpha) x y \\
&= \alpha x^2 + (1-\alpha) y^2 - \alpha(1-\alpha)(x^2 + y^2 - 2xy) \\
&= \alpha x^2 + (1-\alpha) y^2 - \alpha(1-\alpha)(x-y)^2 \\
&\le \alpha x^2 + (1-\alpha) y^2 = \alpha f(x) + (1-\alpha) f(y)
\end{aligned}$$



Example 2: $f(x) = \|x\|^2$

$$\begin{aligned}
f(\alpha x + (1-\alpha) y) &= \|\alpha x + (1-\alpha) y\|^2 = \alpha^2 \|x\|^2 + (1-\alpha)^2 \|y\|^2 + 2\alpha(1-\alpha)\langle x, y\rangle \\
&= \alpha \|x\|^2 + (1-\alpha)\|y\|^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (\alpha^2 - \alpha)\|x\|^2 + (\alpha^2 - \alpha)\|y\|^2 \\
&= \alpha \|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\left(\|x\|^2 + \|y\|^2 - 2\langle x, y\rangle\right) \\
&= \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)\|x-y\|^2 \le \alpha f(x) + (1-\alpha) f(y).
\end{aligned}$$

Example 3: $f(x) = \|x\|$, where $\|x\|$ is a norm on $\mathbb{R}^n$ (e.g., p-norm, 1-norm, ...)

$$f(\alpha x+(1-\alpha)y) = \|\alpha x+(1-\alpha)y\| \le \|\alpha x\|+\|(1-\alpha)y\|$$
$$\le \alpha\|x\|+(1-\alpha)\|y\| \qquad \forall\ \alpha\in[0,1],\quad x,y\in\mathbb{R}^n$$

Therefore, any norm function is convex.

Example 4: Any affine function f is convex, since

$$f(\alpha x+(1-\alpha)y) = \alpha f(x)+(1-\alpha)f(y) \le \alpha f(x)+(1-\alpha)f(y).$$
$$\forall\ \alpha\in[0,1],\ x,y\in\mathbb{R}^n.$$

Example 5: Let $f_1,\cdots,f_n$ are convex, then $f=\sum_{i=1}^n c_i f_i$, $c_i\ge 0$, is convex.

$$f(\alpha x+(1-\alpha)y)=\sum_{i=1}^n c_i f_i(\alpha x+(1-\alpha)y) \le \sum_{i=1}^n c_i\big(\alpha f_i(x)+(1-\alpha)f_i(y)\big)$$
$$=\alpha\sum_{i=1}^n c_i f_i(x)+(1-\alpha)\sum_{i=1}^n c_i f_i(y) = \alpha f(x)+(1-\alpha)f(y)$$

Example 6: Let f be convex and g be affine.

Then $f\circ g$ is convex.

$$(f\circ g)(\alpha x+(1-\alpha)y) = f\big(g(\alpha x+(1-\alpha)y)\big) = f\big(\alpha g(x)+(1-\alpha)g(y)\big)$$
$$\le \alpha f(g(x))+(1-\alpha)f(g(y)).$$

Theorem: If f is convex and differentiable, then

$$x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x) \iff \nabla f(x^{(*)})=0.$$

To prove the theorem, we need a Lemma, which is also useful later.

$$\boxed{\begin{array}{l}\text{Lemma: If } f \text{ is differentiable, then}\\ \qquad f \text{ is convex} \iff f(y)\ge f(x)+\langle\nabla f(x),\, y-x\rangle \quad \forall\ x,y.\end{array}}$$

proof. We prove the one variable case, i.e., $f:\mathbb{R}\to\mathbb{R}$.

"$\Rightarrow$". By convexity, $f(\alpha x+(1-\alpha)y)=f(x+(1-\alpha)(y-x))\le\alpha f(x)+(1-\alpha)f(y)$

This implies $f(y)\ge f(x)+\dfrac{f(x+(1-\alpha)(y-x))-f(x)}{(1-\alpha)(y-x)}(y-x)$

Let $\alpha\to 1$, then $f(y)\ge f(x)+f'(x)\cdot(y-x)$

"$\Leftarrow$" Choose $x\ne y$, and $\alpha\in[0,1]$. Consider $z=\alpha x+(1-\alpha)y$.

Then $f(x)\ge f(z)+f'(z)(x-z)$ (1)

$$f(y) \geq f(z) + f'(z)(y-z) \qquad (2)$$

$$(1)\times\alpha + (2)\times(1-\alpha) \Rightarrow f(z) \leq \alpha f(x) + (1-\alpha) f(y).$$

Next, we prove the multivariable case, i.e, $f: \mathbb{R}^n \to \mathbb{R}$.

"$\Rightarrow$". Consider a function $g(t) = f(tx + (1-t)y)$, $t \in \mathbb{R}$.

So $g'(t) = \langle \nabla f(tx + (1-t)y), x-y \rangle$ by chain rule.

Since f is convex,

$$g(\alpha s + (1-\alpha)t) = f((\alpha s + (1-\alpha)t)x + (1-\alpha s - (1-\alpha)t)y)$$

$$= f(\alpha(sx + (1-s)y) + (1-\alpha)(tx + (1-t)y))$$

$$\leq \alpha f(sx + (1-s)y) + (1-\alpha) f(tx + (1-t)y) = \alpha g(s) + (1-\alpha) g(t)$$

which implies $g(t)$ is convex.

The result from one variable case leads to

$$g(0) \geq g(1) + g'(1)\cdot(-1),$$

i.e., $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

"$\Rightarrow$". Choose $x \neq y$, $\alpha \in [0,1]$, consider $z = \alpha x + (1-\alpha) y$.

Then $f(x) \geq f(z) + \langle \nabla f(z), x-z \rangle \qquad (1)$

$f(y) \geq f(z) + \langle \nabla f(z), y-z \rangle \qquad (2)$

$$(1)\times\alpha + (2)\times(1-\alpha) \Rightarrow f(z) \leq \alpha f(x) + (1-\alpha) f(y). \quad \boxtimes$$

With the lemma, now we can prove the theorem.

proof of the theorem: We prove only $\nabla f(x^{(*)}) = 0 \Rightarrow x^{(*)} = \arg\min_x f(x)$

Since f is convex and differentiable, for any $x \in \mathbb{R}^n$,

$$f(x) \geq f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle$$

By assumption, $\nabla f(x^{(*)}) = 0$

$$\Rightarrow f(x) \geq f(x^{(*)}), \quad \text{i.e., } x^{(*)} = \arg\min f(x). \quad \boxtimes$$

- When the global minimizer is unique?

Strictly convex function: A function $f: \mathbb{R}^n \to \mathbb{R}$ is strictly convex if

$$f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y) \quad \forall\, x \neq y,\ \alpha \in (0,1)$$

- Example 1: $f(x) = x^2$, $x \in \mathbb{R}$.

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)(x-y)^2 \quad \forall\, x, y \in \mathbb{R}^n,\ \alpha \in [0,1].$$

Therefore, $f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y)$ if $x \neq y$, $\alpha \neq 0, \alpha \neq 1$.

Thus, f is strictly convex.

- Example 2: $f(x) = \|x\|_2^2$, $x \in \mathbb{R}^n$, is strictly convex.
- Example 3: $f(x) = \|x\|_1$, $x \in \mathbb{R}^n$, is NOT strictly convex.

Because: Let $x = e_1$, $y = e_2$,

$$f(\alpha x + (1-\alpha) y) = |\alpha| + |1-\alpha| = \alpha + (1-\alpha) = \alpha \|x\|_1 + (1-\alpha) \|y\|_1$$

$$= \alpha f(x) + (1-\alpha) f(y) \quad \forall\, \alpha \in [0,1].$$

- Example 4: If $f_1, f_2, \cdots, f_n$ are strictly convex, then

$f = \sum_{i=1}^{n} c_i f_i$, where $c_i \geq 0$ for all i and not all c_i's are 0,

is strictly convex.

Theorem: Assume f is strictly convex. Then the solution of $\min_{x \in \mathbb{R}^n} f(x)$ is unique if it exists.

proof. Suppose there are at least two solutions. $x^{(*)}, y^{(*)}$. Then $f(x^{(*)}) = f(y^{(*)})$.

Consider $z = \alpha x^{(*)} + (1-\alpha) y^{(*)}$ with $\alpha \in (0,1)$.

Then $z \neq x^{(*)}$, $z \neq y^{(*)}$. Moreover,

$$f(z) = f(\alpha x^{(*)} + (1-\alpha) y^{(*)})$$

$$< \alpha f(x^{(*)}) + (1-\alpha) f(y^{(*)}) = \alpha f(x^{(*)}) + (1-\alpha) f(x^{(*)})$$

$$= f(x^{(*)}),$$

i.e., $f(z) < f(x^{(*)})$, which contradicts with $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x)$.

$\boxtimes$

- Gradient Descent

The simplest algorithm for finding a solution of $\min_{x \in \mathbb{R}^n} f(x)$ is gradient descent in case that f is differentiable.

Let $x^{(k)}$ be the current estimation.

We want to find $x^{(k+1)}$ such that $f(x^{(k+1)}) < f(x^{(k)})$.

Since f is differentiable, we can expand it at $x^{(k)}$ as

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

if $\|x^{(k+1)} - x^{(k)}\|$ is small. This implies

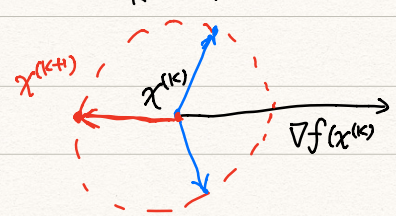
$$f(x^{(k+1)}) - f(x^{(k)}) \approx \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

We want to decrease the value of f the most. So we minimize $f(x^{(k+1)}) - f(x^{(k)})$, i.e, $\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$.

Suppose $\|x^{(k+1)} - x^{(k)}\|$ is fixed. By Cauchy-Schwartz,

$$\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle \geq - \|\nabla f(x^{(k)})\| \|x^{(k+1)} - x^{(k)}\|,$$

and the lower bound is achieved when $x^{(k+1)} - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$ with some positive constant $\alpha_k > 0$.



We obtain

$$\boxed{x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}), \quad k = 0, 1, 2, \cdots}$$

This is called gradient descent algorithm.

- Choice of step size α_k.

To find a proper step size α_k, we need 2-nd order expansion:

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{2} \langle \nabla^2 f(x^{(k)})(x^{(k+1)} - x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

$$= f(x^{(k)}) - \alpha_k \|\nabla f(x^{(k)})\|^2 + \frac{1}{2}\alpha_k^2 \langle \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}), \nabla f(x^{(k)}) \rangle$$

To bound the 3-rd term, we need the assumption that

$$\boxed{\text{Bounded Hessian: } \|\nabla^2 f(x)\| \le M, \quad \forall x \in \mathbb{R}^n}$$

Then,

$$\langle \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}), \nabla f(x^{(k)}) \rangle \le \|\nabla^2 f(x^{(k)}) \nabla f(x^{(k)})\| \cdot \|\nabla f(x^{(k)})\|$$

$$\le \|\nabla^2 f(x^{(k)})\| \cdot \|\nabla f(x^{(k)})\| \cdot \|\nabla f(x^{(k)})\| \le M \|\nabla f(x^{(k)})\|^2.$$

So,

$$f(x^{(k+1)}) \lesssim f(x^{(k)}) - \alpha_k \|\nabla f(x^{(k)})\|^2 + \frac{M}{2}\alpha_k^2 \|\nabla f(x^{(k)})\|^2$$

$$= f(x^{(k)}) - \alpha_k \left(1 - \frac{M}{2}\alpha_k\right) \|\nabla f(x^{(k)})\|^2,$$

which implies

$$f(x^{(k+1)}) - f(x^{(k)}) \lesssim -\alpha_k \left(1 - \frac{M}{2}\alpha_k\right) \|\nabla f(x^{(k)})\|^2. \qquad (**)$$

Therefore, if $\alpha_k(1 - \frac{M}{2}\alpha_k) > 0$, i.e., $0 < \alpha_k < \frac{2}{M}$,

then $f(x^{(k+1)}) - f(x^{(k)}) < 0$.

In other words, $\{f(x^{(k)})\}_k$ is a decreasing sequence.
If we further assume $\boxed{\min_{x \in \mathbb{R}^n} f(x) \text{ has a solution}}$ $\Big\} \Rightarrow \{f(x^{(k)})\}_k$ is bounded below.

Therefore, $\lim_{k \to +\infty} f(x^{(k)}) = c$ for some $c \in \mathbb{R}$.

Furthermore, summing equation $(**)$ over all k yields

$$f(x^{(k+1)}) - f(x^{(0)}) \le -\sum_{i=0}^{k} \alpha_i \left(1 - \frac{M}{2}\alpha_i\right) \|\nabla f(x^{(i)})\|^2$$

If we assume $\boxed{0 < \alpha_l \le \alpha_k \le \alpha_u < \frac{2}{M}}$ for some fixed α_l, α_u for all k,

$$\alpha_k \left(1 - \frac{M}{2}\alpha_k\right) \ge \beta \text{ for some } \beta > 0.$$

It gives

$$\beta \sum_{i=0}^{k} \|\nabla f(x^{(i)})\|^2 \le \sum_{i=0}^{k} \alpha_i \left(1 - \frac{M}{2}\alpha_i\right) \|\nabla f(x^{(i)})\|^2$$

$$\leq f(x^{(0)}) - f(x^{(i+1)})$$

Sending $k \to \infty$, we obtain

$$\sum_{i=0}^{+\infty} \|\nabla f(x^{(i)})\|^2 \leq (f(x^{(0)}) - c)/\beta < +\infty$$

Thus, $\lim_{k \to +\infty} \|\nabla f(x^{(k)})\| = 0$

To sum up,

Theorem: Assume ① f is differentiable and $\exists$ a solution of $\min_{x \in \mathbb{R}^n} f(x)$.

② $\|\nabla^2 f(x)\| \leq M$ for all $x \in \mathbb{R}^n$

③ $0 < \alpha_l \leq \alpha_k \leq \alpha_u < \frac{2}{M}$ for some α_l, α_u and for all k.

Then, the sequence $\{x^{(k)}\}$ generated by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

Satisfies: i). $f(x^{(k+1)}) < f(x^{(k)})$ (the function value decreases)

ii) $\lim_{k \to +\infty} \|\nabla f(x^{(k)})\| = 0$ (the limit has a vanishing gradient).

Since a vanishing gradient is not a sufficient condition for a global minimizer, the gradient descent is NOT guaranteed to find a solution of $\min_{x \in \mathbb{R}^n} f(x)$. It finds only a vanishing gradient with a decreasing function value of f.

In the special case "f is convex", the gradient descent will finds a solution of $\min_{x \in \mathbb{R}^n} f(x)$, as the global minimizer is equivalent to a vanishing gradient.

- Case Studies:

① Least Squares:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2, \quad \text{where } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

Let $f(x) = \frac{1}{2} \|Ax - b\|_2^2$.

- $f(x)$ is convex, because:

Let $f_1(x) = Ax - b$, $f_2(y) = \frac{1}{2}\|y\|_2^2$. Then $f = f_2 \circ f_1$

Since f_1 is affine and f_2 is convex, f is convex.

- Obviously, f is differentiable, and

$$\nabla f(x) = A^T(Ax - b).$$

- Therefore,

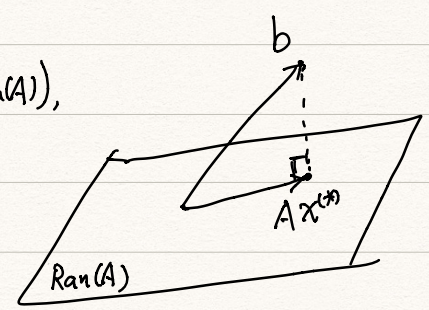
$$x^{(*)} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 \iff \boxed{A^T A x^{(*)} = A^T b}$$

called the Normal equation of Least Squares.

Geometric explanation:

Ax is always in the range of A ($\mathrm{Ran}(A)$),

b is NOT necessarily in $\mathrm{Ran}(A)$



Therefore,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 \iff \min_{y \in \mathrm{Ran}(A)} \frac{1}{2}\|y - b\|_2^2$$

i.e., $Ax^{(*)}$ is the projection of b onto $\mathrm{Ran}(A)$. So,

$$b - Ax^{(*)} \perp \mathrm{Ran}(A),$$

which is the same as

$$\langle Ax^{(*)} - b, Ay \rangle = 0 \quad \forall y \in \mathbb{R}^n.$$

$$\iff \langle A^T(Ax^{(*)} - b), y \rangle = 0 \quad \forall y \in \mathbb{R}^n$$

$$\iff A^T(Ax^{(*)} - b) = 0.$$

- It can be shown that: if $A^T A$ is invertible, then $f(x)$ is strictly convex, and therefore the solution of least squares is unique.
- To find $x^{(*)}$, we may use a gradient descent

$$x^{(k+1)} = x^{(k)} - \alpha_k A^T(Ax^{(k)} - b)$$

To choose a good α_k, we may use "line search", i.e., we set

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f\left(x^{(k)} - \alpha A^T(Ax^{(k)} - b)\right).$$

In other words, α_k is the optimal step size

Let $g(\alpha) = f(x^{(k)} - \alpha A^T(Ax^{(k)} - b))$.

It can be checked $g(\alpha)$ is convex, and therefore

$$g'(\alpha_k) = 0,$$

which gives $$\alpha_k = \frac{\|A^T(Ax^{(k)} - b)\|_2^2}{\|AA^T(Ax^{(k)} - b)\|_2^2}.$$

This leads to the steepest descent algorithm for least squares

$$\boxed{\begin{array}{l} \text{Intialize } x^{(0)} \\ \text{for } k = 0, 1, 2, \cdots \\ \qquad g^{(k)} = A^T(Ax^{(k)} - b) \\ \qquad \alpha_k = \dfrac{\|g^{(k)}\|_2^2}{\|Ag^{(k)}\|_2^2} \\ \qquad x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} \\ \text{end} \end{array}}$$

② Neural Network training

Given $\{x^{(i)}, y_i\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

We want to find a function f such that

$$f(x^{(i)}) \approx y_i, \quad i = 1, 2, \cdots, m.$$

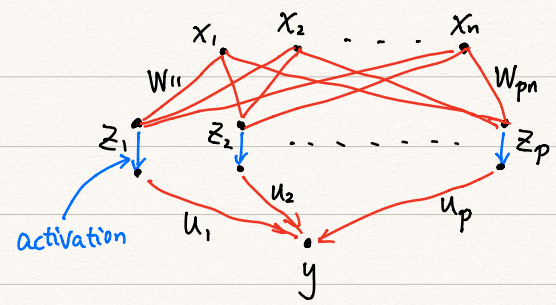
In linear regression, we choose f be an affine function.

In kernel regression, we choose f be a linear function on the feature space.

In deep learning, we choose f be a function generated from an deep neural network.

For simplicity, we consider a fully-connected neural network with one hidden layer

(x_1)



Input layer $\quad x = \begin{pmatrix} \vdots \\ x_n \end{pmatrix} \in \mathbb{R}$

Hidden layer $\quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} \in \mathbb{R}^p$

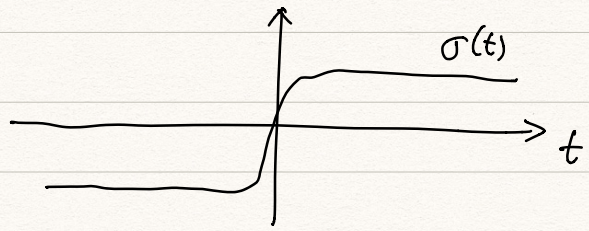
output layer $\quad y \in \mathbb{R}$

Let $W^{(j)} = \begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jn} \end{pmatrix} \in \mathbb{R}^n, \quad u_j \in \mathbb{R}, \quad j = 1, 2, \cdots, p.$

Then, the input-output of the neural network can be written as

$$y = \sum_{j=1}^{p} u_j \, \sigma(\langle W^{(j)}, x \rangle)$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function



Define $\quad f_{W,u}(x) = \sum_{j=1}^{p} u_j \, \sigma(\langle W^{(j)}, x \rangle)$

$$\boxed{\begin{array}{l} \text{Neural network training} \\ \quad \Longleftrightarrow \text{ finding } W, u \text{ s.t. } f_{W,u}(x^{(i)}) \approx y_i \quad i = 1, \cdots, m \end{array}}$$

For this purpose, we minimize

$$\min_{\substack{W \in \mathbb{R}^{p \times n} \\ u \in \mathbb{R}^p}} \sum_{i=1}^{m} \left(f_{W,u}(x^{(i)}) - y_i \right)^2$$

Let $F_i(W,u) = \left(f_{W,u}(x^{(i)}) - y_i \right)^2 = \left(\sum_{j=1}^{p} u_j \, \sigma(\langle W^{(j)}, x^{(i)} \rangle) - y_i \right)^2$

and $\quad F(W,u) = \sum_{i=1}^{m} F_i(W,u).$

We need to solve

$$\min_{W,u} F(W,u)$$

Assume σ is differentiable. Then $F(W,u)$ is differentiable.

So, we may use gradient descent to solve it.

Let's calculate the gradient $\nabla F_i(W,u)$

$\frac{\partial}{\partial u_k} F_i(W,u)$:

$$\frac{\partial}{\partial u_k} F_i(W,u) = \frac{\partial}{\partial u_k}\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)^2$$

$$= 2\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right) \cdot \frac{\partial}{\partial u_k}\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)$$

$$= 2\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right) \cdot \frac{\partial}{\partial u_k}\left(u_k\, \sigma(\langle w^{(k)}, x^{(i)}\rangle)\right)$$

$$= 2\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right) \cdot \sigma(\langle w^{(k)}, x^{(i)}\rangle)$$

$$= 2\left(f_{W,u}(x^{(i)}) - y_i\right) \cdot \sigma(\langle w^{(k)}, x^{(i)}\rangle)$$

$\frac{\partial}{\partial w^{(k)}} F_i(W,u)$:

Let $g(w^{(k)}) = \left(u_k\, \sigma(\langle w^{(k)}, x^{(i)}\rangle) + \underbrace{\sum_{j\neq k} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i}_{=A}\right)^2$

Let $g_1(t) = (t+A)^2 \qquad g_1 : \mathbb{R} \to \mathbb{R}$

$$\Rightarrow Dg_1(t)(s) = 2(t+A)\, s$$

Let $g_2(z) = u_k\, \sigma(z), \qquad g_2 : \mathbb{R} \to \mathbb{R}$

$$\Rightarrow Dg_2(z)(v) = u_k\, \sigma'(z)\, v$$

Let $g_3(w) = \langle w, x^{(i)}\rangle \qquad g_3 : \mathbb{R}^n \to \mathbb{R}$

$$\Rightarrow Dg_3(x)(u) = \langle u, x^{(i)}\rangle$$

Note that $g(w^{(k)}) = g_1(g_2(g_3(w^{(k)})))$

The chain rule gives

$$Dg(w^{(k)})(u) = Dg_1(g_2(g_3(w^{(k)})))\left(Dg_2(g_3(w^{(k)}))\left(Dg_3(w^{(k)})(u)\right)\right)$$

$$= 2\left(u_k\, \sigma(\langle w^{(k)}, x^{(i)}\rangle) + A\right) \cdot u_k\, \sigma'(\langle w^{(k)}, x^{(i)}\rangle) \cdot \langle u, x^{(i)}\rangle$$

$$= 2\left(\sum_{j=1}^{p} u_j\, \sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right) \cdot u_k\, \sigma'(\langle w^{(k)}, x^{(i)}\rangle)\, \langle x^{(i)}, u\rangle$$

$$= 2\left(f_{W,u}(x^{(i)}) - y_i\right) \cdot u_k\, \sigma'(\langle w^{(k)}, x^{(i)}\rangle) \cdot \langle x^{(i)}, u\rangle$$

Therefore,

$$\frac{\partial}{\partial w^{(k)}} F_i(W,u) = 2\left(f_{w,u}(x^{(i)}) - y_i\right) \cdot u_k \, \sigma'(\langle w^{(k)}, x^{(i)} \rangle) \cdot x^{(i)}$$

So, Gradient descent for nueral network training

for $l = 0, 1, 2, \cdots$

$$\begin{cases} u_k^{(l+1)} = u_k^{(l)} - \alpha_l \left(\sum_{i=1}^{m} \frac{\partial}{\partial u_k} F_i(W,u) \right) \\ w^{(k,l+1)} = w^{(k,l)} - \alpha_l \left(\sum_{i=1}^{m} \frac{\partial}{\partial w^{(k)}} F_i(W,u) \right) \end{cases} \quad k = 1, 2, \cdots, p$$

end

Back propogation in deep learning. (Stochastic gradient descent)

for $l = 0, 1, 2, \cdots$

Choose a traing example i_l from $\{1, 2, \cdots, m\}$ randomly.

$$u_k^{(l+1)} = u_k^{(l)} - \alpha_l \frac{\partial}{\partial u_k} F_{i_l}(W,u)$$

$$w^{(k,l+1)} = w^{(k,l)} - \alpha_l \frac{\partial}{\partial w^{(k)}} F_{i_l}(W,u) \quad k = 1, 2, \cdots, p$$

end

§ 4.2 Unconstrained Non-Smooth Convex Optimization

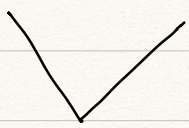
We consider

$$\min_{x \in \mathbb{R}^n} \; g(x)$$

where $g(x)$ is non-differentiable but convex.

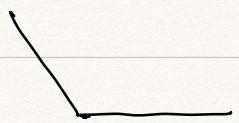
Examples of non-differentiable convex functions.

- 1-norm (in LASSO and sparsity recovery)



- ReLU function (in deep learning)

- loss function $h(t)$ in "soft" SVM



§ 4.2.1. Sub-differential / Sub-gradient. and Optimality

- To give an optimality condition, we need to extend differentiation to non-differentiable convex functions.

  To this end, we first prove:

- Theorem: Assume $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Then, for any given vector $x \in \mathbb{R}^n$,

$$\{\nabla f(x)\} = \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y-x \rangle \quad \forall y \in \mathbb{R}^n\}.$$

proof: Since, if f is convex and differentiable, $\nabla f(x)$ satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall y \in \mathbb{R}^n,$$

we have

$$\{\nabla f(x)\} \subseteq \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y-x \rangle, \; \forall y \in \mathbb{R}^n\}$$

It remains to prove

$$\{\nabla f(x)\} \supseteq \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y-x\rangle, \ \forall y \in \mathbb{R}^n\}$$

To this end, let $u \in \mathbb{R}^n$ be satisfying $f(y) \geq f(x) + \langle u, y-x\rangle \ \forall y \in \mathbb{R}^n$.

Since f is convex, $f(2x-y) \geq f(x) + \langle \nabla f(x), x-y\rangle$

Therefore, $f(y) + f(2x-y) - 2f(x) \geq \langle u - \nabla f(x), y-x\rangle \quad \forall y \in \mathbb{R}^n$. --- (1)

Similarly, $f(2x-y) \geq f(x) + \langle u, x-y\rangle$

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

$\Rightarrow f(y) + f(2x-y) - 2f(x) \geq -\langle u - \nabla f(x), y-x\rangle \quad \forall y \in \mathbb{R}^n$. --- (2)

Since $f(2x-y) + f(y) - 2f(x) = 2\left(f(2x-y) + f(y) - f\left(\frac{1}{2}(2x-y) + \frac{1}{2}y\right)\right) \geq 0$,

combing (1) & (2) gives $|\langle u - \nabla f(x), y-x\rangle| \leq |f(y) + f(2x-y) - 2f(x)|$.

Also, $\left.\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y-x\rangle + o(\|x-y\|) \\ f(2x-y) &= f(x) + \langle \nabla f(x), x-y\rangle + o(\|x-y\|) \end{aligned}\right\} \Rightarrow f(y) + f(2x-y) - 2f(x) = o(\|x-y\|)$,

i.e., $\lim\limits_{y \to x} \dfrac{|f(y) + f(2x-y) - 2f(x)|}{\|x-y\|} = 0$.

which implies

$$0 \leq \lim_{y \to x} \frac{|\langle u - \nabla f(x), y-x\rangle|}{\|y-x\|} \leq \lim_{y \to x} \frac{|f(y) + f(2x-y) - 2f(x)|}{\|x-y\|} = 0$$

$$\Rightarrow \lim_{y \to x} \frac{|\langle u - \nabla f(x), y-x\rangle|}{\|y-x\|} = 0.$$

Now, if we take $y = x + t(u - \nabla f(x))$ with $t \in \mathbb{R}$,

then $\lim\limits_{t \to 0} \dfrac{|\langle u - \nabla f(x), y-x\rangle|}{\|y-x\|} = \lim\limits_{t \to 0} \dfrac{t\|u - \nabla f(x)\|^2}{t\|u - \nabla f(x)\|} = \|u - \nabla f(x)\|$

$$\| $$

$$\lim_{y \to x} \frac{|\langle u - \nabla f(x), y-x\rangle|}{\|y-x\|} = 0$$

Therefore $u = \nabla f(x)$. $\boxtimes$.

- Therefore, we can use the set $\{u \mid f(y) \geq f(x) + \langle u, y-x\rangle, \forall y \in \mathbb{R}^n\}$ to define gradient of a convex differentiable function f.

  Notice that there is no limit in this equivalent definition. So, we can extend it to convex non-differentiable function.

- Sub-differential / Sub-gradient:

Given a convex function $g: \mathbb{R}^n \to \mathbb{R}$ (differentiable or non-differentiable), its sub-differential at $x \in \mathbb{R}^n$, denoted by $\partial g(x)$, is defined by

$$\partial g(x) = \{ u \in \mathbb{R}^n \mid g(y) \geq g(x) + \langle u, y-x \rangle \quad \forall y \in \mathbb{R}^n \}.$$

Any element in $\partial g(x)$ is call a sub-gradient.

- From previous theorem, if g is differentiable at x, then

$$\partial g(x) = \{ \nabla f(x) \}.$$

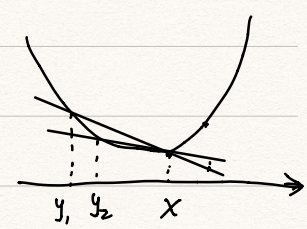
- What if g is not differentiable at $x \in \mathbb{R}^n$?

Theorem: If g is NOT differentiable at $x \in \mathbb{R}^n$, then $\partial g(x)$ contains more than one element.

proof. We prove only the 1-D case, i.e., $g: \mathbb{R} \to \mathbb{R}$ is convex.

Let $y_1 \leq y_2 < x$. Then, $\exists\, t \in (0,1]$ s.t.

$$y_2 = t y_1 + (1-t) x = x + t(y_1 - x)$$



Let $S(y) = \dfrac{g(y) - g(x)}{y - x}$.

Since g is convex,

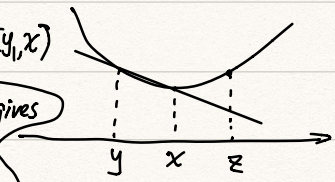
$$\begin{aligned}
S(y_2) - S(y_1) &= \frac{g(y_2) - g(x)}{y_2 - x} - \frac{g(y_1) - g(x)}{y_1 - x} \\
&= \frac{g(t y_1 + (1-t)x) - g(x)}{t y_1 + (1-t)x - x} - \frac{g(y_1) - g(x)}{y_1 - x} \\
&\geq \frac{t g(y_1) + (1-t) g(x) - g(x)}{t(y_1 - x)} - \frac{g(y_1) - g(x)}{y_1 - x} \\
&= \frac{t(g(y_1) - g(x))}{t(y_1 - x)} - \frac{(y_1)\, g(x)}{y_1 - x} = 0
\end{aligned}$$

(because $y_2 - x < 0$)

Therefore, $S(y)$ is monotonically non-increasing on $[y_1, x)$.

Furthermore, let $z > x$. Then, due to the convexity,

(A) $\cdots\cdots$ $g(z) \geq g(x) + S(y)(z - x), \quad \forall y \in [y_1, x)$



because otherwise letting $t = \dfrac{z-x}{z-y} \in (0,1)$ gives

$$x = z + t(y - z) = t y + (1-t) z$$

and

$$\begin{aligned}
(x) &> g(z) - \frac{g(y) - g(x)}{y - x}(z - x) \\
&= g(z) - \frac{g(y) - g(x)}{(t-1)(z-y)} \cdot t(z - y) = g(z) + \frac{t(g(y) - g(x))}{1-t}
\end{aligned}$$

$$\Rightarrow (1-t)g(x) > (1-t)g(z) + t\,g(y) - t\,g(x)$$

$$\Rightarrow g(x) = g(ty+(1-t)z) > (1-t)g(z) + t\,g(y).$$

Contradiction with convexity of g.

Eq. (A) implies $s(y) \leq \dfrac{g(z)-g(x)}{z-x} \equiv s(z)$.

To sum up, $s(y)$ is monotonically non-increasing on $[y_1, x)$

and $s(y) \leq s(z) \quad \forall\, z > x$.

Therefore, $\lim\limits_{y \to x_-} s(y) \equiv D_-g(x)$ exists and

Similarly, $\lim\limits_{z \to x_+} s(z) \equiv D_+g(x)$ exisists.

Also, $s(y) \leq D_-g(x) \leq D_+g(x) \leq s(z) \quad \forall\, y<x$ and $z>x$.

Choose $u \in [D_-g(x), D_+g(x)]$. Then:

Case $y<x$: $\quad g(y) = g(x) + \left(\dfrac{g(y)-g(x)}{y-x}\right)(y-x)$

$$= g(x) + s(y)(y-x) \geq (x) + u(y-x)$$

Case $y>x$: $\quad g(y) = g(x) + \left(\dfrac{g(y)-g(x)}{y-x}\right)(y-x)$

$$= g(x) + s(y)(y-x) \geq (x) + u(y-x)$$

Case $y=x$: $\quad g(y) = g(x) + u(y-x)$

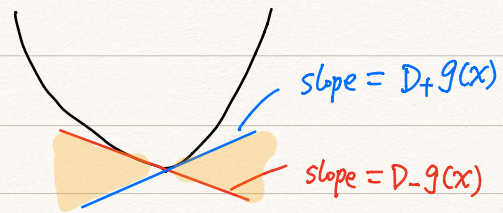
Thus, $\partial g(x) = [D_-g(x), D_+g(x)]$, which is not empty.

When g is differentiable, $D_-g(x) = D_+g(x)$, i.e., there is only one element in $\partial g(x)$.

When g is Not differentiable, $D_-g(x) < D_+g(x)$, $\partial g(x)$ is an interval and contains infinitely many elements. $\boxtimes$.

- From the proof, we see that: the one-sided derivative of a convex function $g: \mathbb{R} \to \mathbb{R}$ always exists. Also,

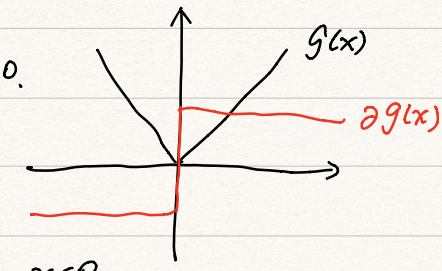
$$\partial g(x) = [D_-g(x), D_+g(x)]. \qquad \forall x \in \mathbb{R}.$$



Example 1: $g(x) = |x| \quad x \in \mathbb{R}$.

$g(x)$ is NOT differentiable at only $x=0$.

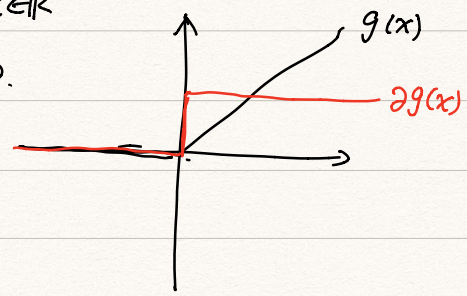
$$\partial g(x) = \begin{cases} \{-1\} & \text{if } x<0 \\ [-1,1] & \text{if } x=0 \\ \{1\} & \text{if } x>0 \end{cases}$$



Example 2: $g(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } x \ge 0 \end{cases}, \quad x \in \mathbb{R}$

$g(x)$ is NOT differentiable at only $x=0$.

$$\partial g(x) = \begin{cases} \{0\} & \text{if } x<0 \\ [0,1] & \text{if } x=0 \\ \{1\} & \text{if } x>0 \end{cases}$$



Example 3: $g(x) = \|x\|_2, \quad x \in \mathbb{R}^n$.

$g(x)$ is differentiable at $x \neq 0$, and $\nabla g(x) = \frac{x}{\|x\|_2}$. if $x \neq 0$.

Therefore $\partial g(x) = \left\{ \frac{x}{\|x\|_2} \right\}$ if $x \neq 0$.

Let us find $\partial g(0)$ by definition.

$$\partial g(0) = \{ u \in \mathbb{R}^n \mid g(y) \ge g(0) + \langle u, y \rangle, \ \forall y \in \mathbb{R}^n \}$$

$$= \{ u \in \mathbb{R}^n \mid \|y\|_2 \ge \langle u, y \rangle, \ \forall y \in \mathbb{R}^n \} \equiv S$$

Claim: $S = U \equiv \{ u \in \mathbb{R}^n \mid \|u\|_2 \le 1 \}$.

Proof. For any $u \in U$, $\|u\|_2 \le 1$, which implies

$$\langle u, y \rangle \le \|u\|_2 \|y\|_2 \le \|y\|_2, \ \forall y \in \mathbb{R}^n \implies u \in S.$$

So. $U \subseteq S$.

For any $u \notin U$, i.e., $\|u\|_2 > 1$, we choose $y = u$.

Then $\langle u, y \rangle = \|u\|_2^2 = \|u\|_2 \|y\|_2 > \|y\|_2$

So $u \notin S$. Thus, $U^c \subseteq S^c \implies S \subseteq U$.

Therefore $S = U$. $\boxtimes$

Altogether, $$\partial \|x\|_2 = \begin{cases} \left\{ \frac{x}{\|x\|_2} \right\} & \text{if } x \neq 0 \\ \{ u \in \mathbb{R}^n \mid \|u\|_2 \leq 1 \} & \text{if } x = 0. \end{cases}$$

- Sub-differential calculus rules:
  - $\partial(\alpha g) = \alpha \partial g \qquad \forall \alpha \in \mathbb{R}$.
  - $\partial(g_1 + g_2) = \partial g_1 + \partial g_2$
  - If $g(x) = f(Ax+b)$, then $\partial g(x) = A^T \partial f(Ax+b)$

Example 4: If $g(x) = \sum_{i=1}^n g_i(x_i)$, where $g_i : \mathbb{R} \to \mathbb{R}$ is convex,

then $$\partial g(x) = \left\{ \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \middle| \; u_i \in \partial g_i(x_i) \right\}$$

proof. Let $\tilde{g}_i(x) = g_i(x_i)$.

Then $\partial g(x) = \partial\left(\sum_{i=1}^n \tilde{g}_i(x)\right) = \sum_{i=1}^n \partial \tilde{g}_i(x)$.

Now lets show $\partial \tilde{g}_i(x) = \{ a e_i \mid a \in \partial g_i(x_i) \}$. by direct calculation

$$\partial \tilde{g}_i(x) = \{ u \in \mathbb{R}^n \mid \tilde{g}_i(y) \geq \tilde{g}_i(x) + \langle u, y-x \rangle, \; \forall y \in \mathbb{R}^n \}$$
$$= \{ u \in \mathbb{R}^n \mid g_i(y_i) \geq g_i(x_i) + u_i \cdot (y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j), \; \forall y \in \mathbb{R}^n \}$$

- $\partial \tilde{g}_i(x) \supseteq \{ u \in \mathbb{R}^n \mid u_i \in \partial g_i(x_i), \; u_j = 0 \; \forall j \neq i \} = \{ a e_i \mid a \in \partial g_i(x_i) \}$
- If $u \neq a e_i$, where $a \in \partial g_i(x_i)$, and $u \in \partial \tilde{g}_i(x)$, then,
  - If $u_j = 0 \; \forall j \neq i$, then

$$g_i(y_i) \geq g_i(x_i) + u_i(y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j) = g_i(x_i) + u_i(y_i - x_i) \quad \forall y_i \in \mathbb{R}$$
$$\Rightarrow u_i \in \partial g_i(x_i) \Rightarrow u = a e_i, \; a \in \partial g_i(x_i)$$

  - Therefore, $\exists j \neq i$ s.t $u_j \neq 0$. Choose $y = (c|u_j| + x_j) e_j$ with $c \in \mathbb{R}$

$$g_i(0) \geq g_i(x_i) + u_i(y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j)$$
$$= g_i(x_i) + c|u_j|^2 \longrightarrow +\infty \quad \text{as} \quad c \to +\infty. \quad \text{contradiction.}$$

Thus, $\partial \tilde{g}_i(x) = \{ a e_i \mid a \in \partial g_i(x_i) \}$

We obtain $\partial g(x) = \{ \sum_{i=1}^n a_i e_i \mid a_i \in \partial g_i(x_i) \}$

$$= \left\{ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \middle| \; u_i \in \partial g_i(x_i) \right\}$$

Example 5: $\partial \|x\|_1 = \left\{ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \middle| \; u_i \in \partial |x_i| \right\}$, because $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

Example 6: $\partial \|Dx\|_1 = \left\{ D^T \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \middle| \; u_i \in \partial |(Dx)_i|_1 \right\}$, where $D \in \mathbb{R}^{m \times n}$.

- Optimality:

$$\boxed{\begin{array}{l} \text{Fermat's Lemma: Assume } g: \mathbb{R}^n \to \mathbb{R} \text{ is convex. Then} \\ \qquad x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff 0 \in \partial g(x^{(*)}) \end{array}}$$

Proof. $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff g(x) \geq g(x^{(*)}) \quad \forall \; x \in \mathbb{R}^n$

$\iff g(x) \geq g(x^{(*)}) + \langle 0, x - x^{(*)} \rangle \quad \forall \; x \in \mathbb{R}^n$

$\iff 0 \in \partial g(x^{(*)})$ $\boxtimes$

Example 1: If g is convex and differentiable, then $\partial g(x) = \{ \nabla g(x) \}$.

Therefore, $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff \nabla g(x^{(*)}) = 0$.

Example 2: $g(x) = \|x\|_1$

$$0 \in \partial \|x^{(*)}\|_1 \iff 0 \in \left\{ \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \middle| \; u_i \in \partial |x_i^{(*)}| \right\} \iff x^{(*)} = 0.$$

Therefore, $g(x) = \|x\|_1$ is minimized at 0.

Example 3: $g(x) = \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

$$\partial g(x) = A^T(Ax - b) + \lambda \, \partial \|x\|_1$$

Therefore, $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff 0 \in A^T(Ax^{(*)} - b) + \lambda \, \partial \|x^{(*)}\|_1$

$\iff \frac{1}{\lambda} A^T(b - Ax^{(*)}) \in \partial \|x^{(*)}\|_1$

Therefore, $\left[\frac{1}{\lambda} A^T(b - Ax^{(*)}) \right]_i \in [-1, 1]$.

If $\left| \left[\frac{1}{\lambda} A^T(b - Ax^{(*)}) \right]_i \right| < 1$, then $[x^{(*)}]_i = 0$.

§ 4.2.2. Sub-Gradient Descent

Consider $\min_{x \in \mathbb{R}^n} g(x)$, where g is convex but non-differentiable.

- Given $x^{(k)} \in \mathbb{R}^n$, let $u^{(k)} \in \partial g(x^{(k)})$. Then, by convexity,

$$g(x^{(k+1)}) \geq g(x^{(k)}) + \langle u^{(k)}, x^{(k+1)} - x^{(k)} \rangle$$

$$\Rightarrow g(x^{(k+1)}) - g(x^{(k)}) \geq \langle u^{(k)}, x^{(k+1)} - x^{(k)} \rangle$$

The right-hand side is minimized if

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k)}, \quad \text{where } \alpha_k > 0.$$

This is called sub-gradient descent.

- However, the right-hand side is only a lower bound, and there might be a gap as big as $O(\|x^{(k+1)} - x^{(k)}\|)$, which makes $g(x^{(k+1)}) > g(x^{(k)})$ for any $\alpha_k > 0$.

Example: $g(x) = |x|$, $x^{(k)} = 0$. Choose $u^{(k)} = -1 \in \partial g(x^{(k)})$.

Then $x^{(k+1)} = x^{(k)} - \alpha_k(-1) = \alpha_k$.

Left hand side $= g(x^{(k+1)}) - g(x^{(k)}) = \alpha_k$

right hand side $= \langle u^{(k)}, x^{(k+1)} - x^{(k)} \rangle = \langle -1, \alpha_k \rangle = -\alpha_k$.

the gap is

$$g(x^{(k+1)}) - g(x^{(k)}) - \langle u^{(k)}, x^{(k+1)} - x^{(k)} \rangle = 2\alpha_k$$

which is $2\|x^{(k+1)} - x^{(k)}\|. \sim O(\|x^{(k+1)} - x^{(k)}\|)$,

Consequently,

$$g(x^{(k+1)}) - g(x^{(k)}) = \langle u^{(k)}, x^{(k+1)} - x^{(k)} \rangle + 2\alpha_k$$

$$= -\alpha_k \|u^{(k)}\|^2 + 2\alpha_k = \alpha_k > 0.$$

In general, to make sub-gradient work, we have to choose

$$\alpha_k \to 0, \quad \text{as } k \to +\infty.$$

In summary, sub-gradient converges very slowly.

- We may use backward sub-gradient descent.

Let $x^{(k)} \in \mathbb{R}^n$ be current iteration, and $x^{(k+1)} \in \mathbb{R}^n$ be the next iteration.

Instead lower bound of $g(x^{(k+1)}) - g(x^{(k)})$, we estimate an upper bound of $g(x^{(k+1)}) - g(x^{(k)})$, i.e.,

$$g(x^{(k+1)}) - g(x^{(k)}) \leq \cdots\cdots$$

So we need $\quad g(x^{(k)}) \geq g(x^{(k+1)}) \cdots\cdots$

The idea is to expand g at $x^{(k+1)}$. For this, we need sub-gradient at $x^{(k+1)}$. Let $u^{(k+1)} \in \partial g(x^{(k+1)})$. Then, the convexity implies

$$g(x^{(k)}) \geq g(x^{(k+1)}) + \langle u^{(k+1)}, x^{(k)} - x^{(k+1)} \rangle,$$

i.e., $$g(x^{(k+1)}) - g(x^{(k)}) \leq \langle u^{(k+1)}, x^{(k+1)} - x^{(k)} \rangle \quad \cdots\cdots \quad (A)$$

The right handside is minimized when $x^{(k+1)} - x^{(k)} = -\alpha_k u^{(k+1)}$, i.e.,

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k+1)}, \qquad \alpha_k \geq 0 \quad \cdots\cdots \quad ($$

This is called backward sub-gradient descent.

Then, (A) implies

$$g(x^{(k+1)}) - g(x^{(k)}) \leq -\alpha_k \|u^{(k+1)}\|^2 \quad \cdots\cdots \quad (B)$$

Thus,

(i). if $\alpha_k > 0$, then $g(x^{(k+1)}) \leq g(x^{(k)})$, i.e., the backward sub-gradient descent gives a monotonical non-increasing $\{g(x^{(k)})\}_{k=0}^{+\infty}$.

Moreover, if $\exists$ a solution of $\min_{x \in \mathbb{R}^n} g(x)$, (i.e. $\exists\, x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x)$), then $\lim_{k \to +\infty} g(x^{(k)}) = C$ exists.

(ii). if $\alpha_k \geq \alpha > 0$, then summing (B) over k gives

$$g(x^{(k+1)}) - g(x^{(0)}) \leq -\sum_{k=0}^{k} \alpha_k \|u^{(k+1)}\|^2 \leq -\alpha \sum_{k=0}^{k} \|u^{(k+1)}\|^2$$

Sending $k \to +\infty$,

$$C - g(x^{(0)}) \leq -\alpha \sum_{k=1}^{\infty} \|u^{(k)}\|^2, \quad \text{i.e.,}$$

$$\sum_{k=1}^{\infty} \|u^{(k)}\|^2 \leq \frac{g(x^{(0)}) - C}{\alpha} < +\infty$$

Therefore, $\lim_{k \to +\infty} \|u^{(k)}\| = 0$, where $u^{(k)} \in \partial g(x^{(k)})$.

If $\{x^{(k)}\}_{k=1}^{+\infty}$ is bounded, i.e., $\exists M>0$ s.t. $\|x^{(k)}\|_2 \le M \ \forall k$,

then $g(x^{(*)}) \ge g(x^{(k)}) + \langle u^{(k)}, x^{(k)} - x^{(*)} \rangle \ge g(x^{(k)}) - \|u^{(k)}\| \, \|x^{(k)} - x^{(*)}\|$

$$\ge g(x^{(k)}) - (M + \|x^{(*)}\|_2) \, \|u^{(k)}\|$$

Sending $k \to +\infty$, we obtain

$\min_{x \in \mathbb{R}^n} g(x) = g(x^{(*)}) \ge C$, which implies $C = \min_{x \in \mathbb{R}^n} g(x)$.

Therefore, $\lim_{k \to +\infty} g(x^{(k)}) = \min_{x \in \mathbb{R}^n} g(x)$.

i.e., backward sub-gradient converges to a global minimum as long as $\alpha_k \ge \alpha > 0$.

Theorem: Consider $\min_{x \in \mathbb{R}^n} g(x)$, where $g: \mathbb{R}^n \to \mathbb{R}$ is convex. Let $\{x^{(k)}\}_k$ be generated by

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k+1)}, \quad \text{where } u^{(k+1)} \in \partial g(x^{(k+1)}).$$

Then, if $\alpha_k \ge \alpha > 0 \ \forall k$ and $\min_{x \in \mathbb{R}^n} g(x)$ has a solution, then,

(i) $\{g(x^{(k)})\}_k$ is monotonically non-increasing;

(ii) $\lim_{k \to +\infty} \|u^{(k)}\|_2 = 0$;

(iii) If $\|x^{(k)}\|_2 \le M \ \forall k$, then $\lim_{k \to +\infty} g(x^{(k)}) = \min_{x \in \mathbb{R}^n} g(x)$.

- However, given $x^{(k)} \in \mathbb{R}^n$, we cannot obtain $x^{(k+1)}$ directly by the iteration:

$$x^{(k+1)} = \underbrace{x^{(k)} - \alpha_k u^{(k+1)}}_{\text{depends on } x^{(k+1)}}, \quad \text{where } u^{(k+1)} \in \partial g(x^{(k+1)})$$

We need to solve the above equation, which is equivalent to

$$x^{(k+1)} \in x^{(k)} - \alpha_k \partial g(x^{(k+1)})$$

$$\Updownarrow$$

$$0 \in x^{(k+1)} - x^{(k)} + \alpha_k \partial g(x^{(k+1)})$$

$$\Updownarrow \leftarrow \text{Fermat's lemma}$$

$$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x), \quad \text{where } F_{\alpha_k}(x) = \frac{1}{2}\|x - x^{(k)}\|_2^2 + \alpha_k g(x)$$

- F_{α_k} is continuous (because all convex function is continuous).
- F_{α_k} is coercive if $\min_{x\in\mathbb{R}^n} g(x)$ exists a solution.

Therefore, $\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ exists at least a solution.

- Furthermore, F_{α_k} is strictly convex.

Therefore, $\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ has a unique solution.

Altogether, $x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ is well-defined.

Thus, backward sub-gradient descent is rewritten as

$$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} \left(\frac{1}{2}\|x - x^{(k)}\|_2^2 + \alpha_k g(x)\right), \quad k = 0, 1, 2, \cdots$$

- Proximity Operator

Definition: Let $g: \mathbb{R}^n \to \mathbb{R}$ be convex. Let $\lambda > 0$ be a parameter.

The proximity operator $\text{prox}_{\lambda g}: \mathbb{R}^n \to \mathbb{R}^n$ of g is defined by

$$\text{prox}_{\lambda g}(y) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x - y\|_2^2 + \lambda g(x), \quad \forall y \in \mathbb{R}^n \qquad \boxtimes$$

- Then, the backward sub-gradient descent is

$$x^{(k+1)} = \text{prox}_{\alpha_k g}(x^{(k)}), \quad k = 0, 1, 2, \cdots$$

For this reason, the backward sub-gradient descent is also called the proximal algorithm.

- Example 1: $g(x) = \|x\|_1$

Then, $x \equiv \text{prox}_{\lambda g}(y) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x - y\|_2^2 + \lambda\|x\|_1$

$$\Updownarrow$$

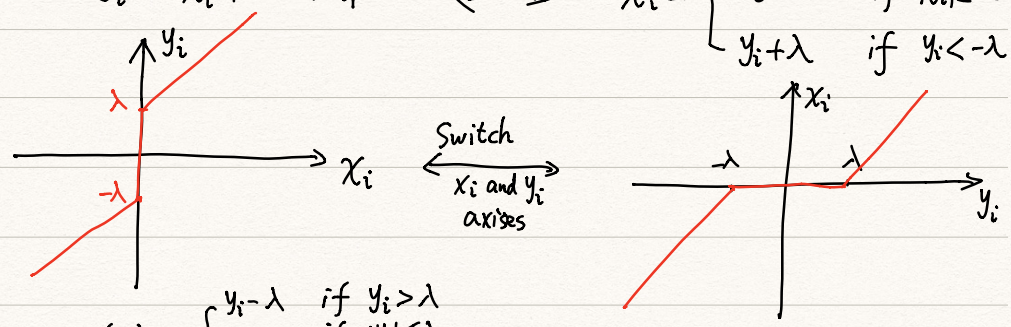
$$0 \in x - y + \lambda \partial\|x\|_1$$

$$\Updownarrow \qquad \left(\text{Recall } \partial\|x\|_1 = \left\{\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \middle| u_i \in \partial|x_i|\right\}\right)$$

$$0 \in x_i - y_i + \lambda \partial|x_i|, \quad i = 1, 2, \cdots, n.$$

$$\Updownarrow$$

$$y_i \in x_i + \lambda \partial |x_i| \iff x_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } |y_i| \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$



We define $T_\lambda(y_i) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } |y_i| \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$

Then, $\operatorname{prox}_{\lambda \|\cdot\|_1}(y) = T_\lambda(y) \equiv \begin{bmatrix} T_\lambda(y_1) \\ T_\lambda(y_2) \\ \vdots \\ T_\lambda(y_n) \end{bmatrix}$ — called soft-thresholding operator.

The proximal algorithm for $\min_{x \in \mathbb{R}^n} \|x\|_1$ gives

$$x^{(k+1)} = T_{\alpha_k}(x^{(k)}), \quad k = 0, 1, 2, \cdots$$

At each iteration, components of $x^{(k)}$ are shrinked by α_k until it becomes 0. Therefore, $x^{(k)} \to 0 \equiv \arg\min_{x \in \mathbb{R}^n} \|x\|_1$. ☒

- Example 2: $g(x) = \|x\|_2^2$

Then $x \equiv \operatorname{prox}_{\lambda g}(y) = \arg\min_{x \in \mathbb{R}^n} \left(\frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_2^2 \right)$

$$\Updownarrow$$

$$x - y + 2\lambda x = 0$$

$$\Updownarrow$$

$$x = \frac{1}{1+2\lambda} y$$

The proximal algorithm for $\min_{x \in \mathbb{R}^n} \|x\|_2^2$ is

$$x^{(k+1)} = \frac{1}{2+2\alpha_k} x^{(k)},$$

which obviously converges to 0, the arg min of $\|x\|_2^2$.

- Example 3: $g(x) = \|x\|_2$

Then $x \equiv \operatorname{prox}_{\lambda g}(y) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_2$

$$\Updownarrow$$

$$0 \in x - y + \lambda \partial \|x\|_2$$

$$\Updownarrow$$

$$y \in x + \lambda \partial \|x\|_2 \quad \cdots\cdots \quad (**)$$

Let us find x.

If $x \neq 0$, then $y = x + \lambda \frac{x}{\|x\|_2} \Rightarrow x = cy$ with $c = \frac{1}{1+\frac{\lambda}{\|x\|_2}} \geq 0$.

If $x = 0$, then $\Longrightarrow$ $x = cy$ with $c = 0$.

In any cases, the solution $x = cy$ with $c \geq 0$.

So the orginal problem $\Longleftrightarrow$ $\min\limits_{c \geq 0} \frac{1}{2}(1-c)^2 \|y\|_2^2 + \lambda |c| \, \|y\|_2$

If $\|y\|_2 = 0$, obviously $c = 0$.

If $\|y\|_2 > 0$, then $\min\limits_{c \geq 0} \frac{\|y\|_2}{2}(1-c)^2 + \lambda |c|$

$$\Longleftrightarrow \min_{c \geq 0} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$$

Since $T_{\frac{\lambda}{\|y\|_2}}(1) \geq 0$ is the solution of $\min\limits_{c \in \mathbb{R}} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$,

it is also the solution $\min\limits_{c \geq 0} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$.

Therefore,
$$c = T_{\frac{\lambda}{\|y\|_2}}(1) = \begin{cases} 1 - \frac{\lambda}{\|y\|_2} & \text{if } 1 \geq \frac{\lambda}{\|y\|_2}. \\ 0 & \text{if } 1 \leq \frac{\lambda}{\|y\|_2}. \end{cases}$$

$$= \begin{cases} \frac{\|y\|_2 - \lambda}{\|y\|_2} & \text{if } \|y\|_2 \geq \lambda \\ 0 & \text{if } \|y\|_2 \leq \lambda. \end{cases}$$

Finally,
$$\text{prox}_{\lambda g}(y) = x = \begin{cases} \frac{\|y\|_2 - \lambda}{\|y\|_2} y & \text{if } \|y\|_2 \geq \lambda \\ 0 & \text{if } \|y\|_2 \leq \lambda. \end{cases}$$

The proximal algorithm for $\min\limits_{x \in \mathbb{R}^n} \|x\|_2$ is

$$x^{(k+1)} = \begin{cases} \frac{\|x^{(k)}\|_2 - \alpha_k}{\|x^{(k)}\|_2} x^{(k)} & \text{if } \|x^{(k)}\|_2 \geq \alpha_k \\ 0 & \text{if } \|x^{(k)}\|_2 \leq \alpha_k \end{cases},$$

which obvious converges to 0.

- In general, it is difficult to find $\text{prox}_{\lambda g}(\cdot)$ for a given convex

function g. Therefore, the proximal algorithm (Backward sub-gradient descent) is NOT practical.

§ 4.2.3. Case Study: LASSO regression

LASSO regression model:

$$\min_{x\in\mathbb{R}^n} \underbrace{\tfrac{1}{2}\|Ax-b\|_2^2}_{\text{fit the given data.}} + \underbrace{\lambda\|x\|_1}_{\text{promote sparsity of the solution}} \qquad \text{(LASSO)}$$

- The objective function is convex.
- The objective function is continuous

  The objective function is coercive, because

  when $\|x\|_2 \to +\infty$, $\quad \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \geq \lambda\|x\|_1 \geq \lambda\|x\|_2 \to +\infty$.

  Therefore, there exists at least one solution.
- The uniqueness of the solution is NOT guaranteed, unless A satisfies some assumptions (e.g. A is invertible)
- The solution of (LASSO) is $\boxed{\text{sparse.}}$

  To see this, let $x\in\mathbb{R}^n$ be a solution. Then

$$0 \in A^T(Ax-b) + \lambda\,\partial\|x\|_1$$

$$\Updownarrow$$

$$0 \in \alpha A^T(Ax-b) + \lambda\alpha\,\partial\|x\|_1 \qquad \forall \alpha>0$$

$$\Updownarrow$$

$$0 \in x - (x-\alpha A^T(Ax-b)) + \lambda\alpha\,\partial\|x\|_1 \qquad \forall \alpha>0$$

$$\Updownarrow$$

$$x = \arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|x-y\|_2^2 + \lambda\alpha\,\partial\|x\|_1, \quad \text{where } y = x-\alpha A^T(Ax-b)$$

$$\Updownarrow$$

$$x = \mathrm{prox}_{\lambda\alpha\|\cdot\|_1}\left(x - \alpha A^T(Ax-b)\right)$$

$$\Updownarrow$$

$$x = T_{\lambda\alpha}\left(x - \alpha A^T(Ax-b)\right)$$

Recall that $T_{\lambda\alpha}(y)$ set y_i to 0 if $|y_i| \le \lambda\alpha$.

Therefore, x is a sparse if λ is large enough.

- Numerical solver for (LASSO).

The objective function in (LASSO) is NON-SMOOTH.

The gradient descent can not be applied.

But if the backward sub-gradient (proximal algorithm) is applied, then we need the proximity operator, which solves

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \alpha\left(\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1\right)$$

This is the same difficulty as (LASSO).

We will use a "mixed" forward and backward (sub-)gradient algorithm.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ — convex and smooth

$g(x) = \|x\|_1$ — Convex but non-smooth

$$\text{(LASSO)} \iff \min_{x\in\mathbb{R}^n} \underbrace{f(x)}_{\text{forward gradient}} + \underbrace{\lambda g(x)}_{\text{backward sub-gradient}}$$

We use the algorithm

$$x^{(k+1)} = x^{(k)} - \alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right), \quad \text{with } u^{(k+1)} \in \partial g(x^{(k+1)})$$

It is rewritten as

$$x^{(k+1)} \in x^{(k)} - \alpha_k \nabla f(x^{(k)}) - \alpha_k \lambda \partial g(x^{(k+1)})$$

$$0 \in x^{(k+1)} - \left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right) + \alpha_k \lambda \partial g(x^{(k+1)}) \quad \cdots\cdots (A)$$

Recall that

$$x = \mathrm{prox}_{\beta g}(y) \iff x = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \beta g(x)$$

$$\iff 0 \in x - y + \beta\partial g(x) \quad \cdots\cdots (B)$$

By comparing (A) and (B), we obtain

$$x^{(k+1)} = \text{prox}_{\alpha_k \lambda g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)$$

Since $\nabla f(x) = \nabla\left(\frac{1}{2}\|Ax-b\|_2^2\right) = A^T(Ax-b)$

$$\text{prox}_{\alpha_k \lambda g}(x) = T_{\alpha_k \lambda}(x),$$

the algorithm is

$$\boxed{x^{(k+1)} = T_{\alpha_k \lambda}\left(x^{(k)} - \alpha_k A^T(Ax^{(k)}-b)\right)}$$

This algorithm is known as iterative soft-thresholding algorithm

- From the derivation, for a generic optimization

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x),$$

where $f(x)$ is convex and smooth

$g(x)$ is convex but non-smooth,

we can use the forward-backward splitting (FBS) algorithm

$$x^{(k+1)} = x^{(k)} - \alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right), \quad u^{(k+1)} \in \partial g(x^{(k+1)})$$

Which is the same as

$$x^{(k+1)} = \text{prox}_{\lambda \alpha_k g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)$$

This is also known as proximal gradient algorithm.

- Convergence

Since f is convex and smooth,

$$(C) \cdots f(x^{(k+1)}) = f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{2}\langle \nabla^2 f(\tilde{x}^{(k)})(x^{(k+1)} - x^{(k)}), x^{(k+1)} - x^{(k)} \rangle,$$

1 Since g is convex,

$$g(x^{(k)}) \geq g(x^{(k+1)}) + \langle u^{(k+1)}, x^{(k)} - x^{(k+1)} \rangle, \quad u^{(k+1)} \in \partial g(x^{(k+1)})$$

which is equivalent to

$$(D) \cdots g(x^{(k+1)}) \leq g(x^{(k)}) + \langle u^{(k+1)}, x^{(k+1)} - x^{(k)} \rangle$$

$(C) + \lambda(D)$ gives

$$\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right]$$

$$\le \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] + \langle \nabla f(x^{(k)}) + \lambda u^{(k+1)}, x^{(k+1)} - x^{(k)} \rangle$$
$$+ \frac{1}{2} \langle \nabla^2 f(\tilde{x}^{(k)})(x^{(k+1)} - x^{(k)}), x^{(k+1)} - x^{(k)} \rangle,$$

Assuming $\langle \nabla^2 f(x)\, u, u \rangle \le M \|u\|_2^2 \quad \forall\, x, u \in \mathbb{R}^n$.

$$\le \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] - \alpha_k \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 + \frac{M\alpha_k^2}{2} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$
$$= \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] - \alpha_k \left(1 - \frac{M\alpha_k}{2}\right) \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$

Assume $0 < \alpha_l \le \alpha_k \le \alpha_u < \frac{2}{M}$ for some $\alpha_l, \alpha_u \in \mathbb{R}$.

Then, $\exists\, \beta$ s.t. $\alpha_k \left(1 - \frac{M}{2}\alpha_k\right) \ge \beta > 0$.

So, $\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right] - \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] \le -\beta \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 \cdots (C)$

$\boxed{\{f(x^{(k)}) + \lambda g(x^{(k)})\}_k \text{ is a monotonically non-increasing sequence.}}$

(2) Assume $\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x)$ has a solution, i.e.,

$\exists\, x^{(*)}$, s.t. $f(x^{(*)}) + \lambda g(x^{(*)}) = \min_{x \in \mathbb{R}^n} f(x) + \lambda g(x)$.

Then, $\lim_{k \to \infty} f(x^{(k)}) + \lambda g(x^{(k)}) = C$ exists.

Furthermore, summing (C) over k, we obtain

$$\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right] - \left[f(x^{(0)}) + \lambda g(x^{(0)})\right] \le -\beta \sum_{k=0}^{k} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$

So $\sum_{k=0}^{k} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 \le \frac{1}{\beta}\left[\left(f(x^{(0)}) + \lambda g(x^{(0)})\right) - \left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right]\right]$

$k \to +\infty$ gives $\sum_{k=0}^{+\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 < +\infty$

Consequently, $\lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 = 0$

Since $\alpha_k \left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right) = x^{(k)} - x^{(k+1)}$

$$0 \le \lim_{k \to +\infty} \|x^{(k)} - x^{(k+1)}\|_2 = \lim_{k \to +\infty} \|\alpha_k \left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right)\|_2$$
$$\le \alpha_u \cdot \lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 = 0,$$

which implies $\lim_{k \to +\infty} \|x^{(k)} - x^{(k+1)}\|_2 = 0$.

This together with the continuity of $\nabla f(x)$ implies $0 = \lim_{k \to +\infty} \|\nabla f(x^{(k)}) - \nabla f(x^{(k+1)})\|_2$

Therefore,

$$0 \le \lim_{k \to +\infty} \|\nabla f(x^{(k+1)}) + \lambda u^{(k+1)}\|_2 \le \lim_{k \to +\infty} \left(\|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 + \|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\|_2\right) = 0,$$

i.e., $\boxed{\lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k)}\|_2 = 0}$

③ Moreover,

$$[f(x^{(*)})+\lambda g(x^{(*)})] \geq [f(x^{(k)})+\lambda g(x^{(k)})] + \langle \nabla f(x^{(k)})+\lambda u^{(k)},\ x^{(*)}-x^{(k)}\rangle$$

$$\geq [f(x^{(k)})+\lambda g(x^{(k)})] - \|\nabla f(x^{(k)})+\lambda u^{(k)}\|_2 \cdot (\|x^{(*)}\|_2+\|x^{(k)}\|_2)$$

Assume $\exists\, B>0$ s.t. $\|x^{(k)}\|_2 \leq B \quad \forall k$.

Then, sending $k\to+\infty$, we obtain

$$\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x) \geq C. \quad \left(\text{Obviously, } C \geq \min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)\right).$$

Thus, $\boxed{\lim_{k\to+\infty} f(x^{(k)})+\lambda g(x^{(k)}) = \min_{k\to+\infty} f(x)+\lambda g(x)}$

To sum up,

Theorem: Consider $\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)$, where $\lambda>0$, f, g are convex.

Assume: ① $\langle \nabla^2 f(x) u, u\rangle \leq M\|u\|_2^2$, $\forall\, x, u\in\mathbb{R}^n$. (i.e., f is smooth and has a bounded Hessian)

② $0<\alpha_l\leq\alpha_k\leq\alpha_u<\frac{2}{M}$ for some $\alpha_l, \alpha_u\in\mathbb{R}$.

③ $\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)$ has a solution.

Then, the sequence $\{x^{(k)}\}_k$ generated by

$$x^{(k+1)} = \mathrm{prox}_{\lambda\alpha_k g}(x^{(k)} - \alpha_k \nabla f(x^{(k)}))$$

satisfies:

(i) $\{f(x^{(k)})+\lambda g(x^{(k)})\}_k$ is monotonically non-increasing,

(ii) $\exists\, u^{(k)}\in\partial g(x^{(k)})$, s.t. $\lim_{x\to+\infty}\|\nabla f(x^{(k)})+u^{(k)}\|_2=0$.

(iii) If $\exists\, B>0$ s.t. $\|x^{(k)}\|_2\leq B \quad \forall k$, then

$$\lim_{k\to+\infty} f(x^{(k)})+\lambda g(x^{(k)}) = \min_{x\in\mathbb{R}^n} f(x)+\lambda g(x).$$

Remark: The condition $\|x^{(k)}\|_2\leq B$ is satisfied if $f(x)+\lambda g(x)$ is coercive. (Why?).

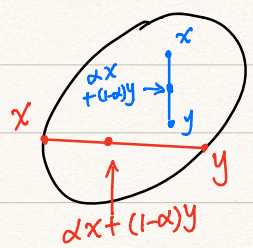
§ 4.3. Constrained Optimization

$$\min_{x \in S} f(x)$$

§ 4.3.1. Convex sets

A set $S \in \mathbb{R}^n$ is convex if

$$x, y \in S \implies \alpha x + (1-\alpha) y \in S \quad \forall\ 0 \le \alpha \le 1$$

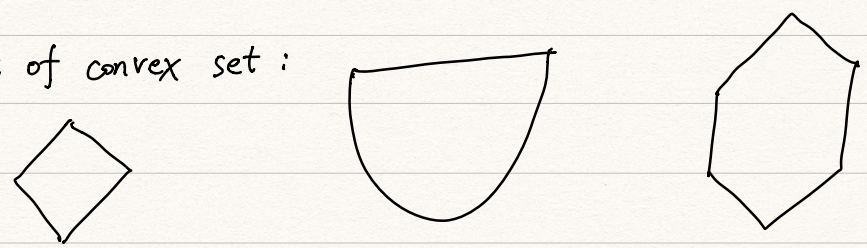


Since $\{\alpha x + (1-\alpha) y \mid 0 \le \alpha \le 1\}$ is the line segment $\overline{xy}$,

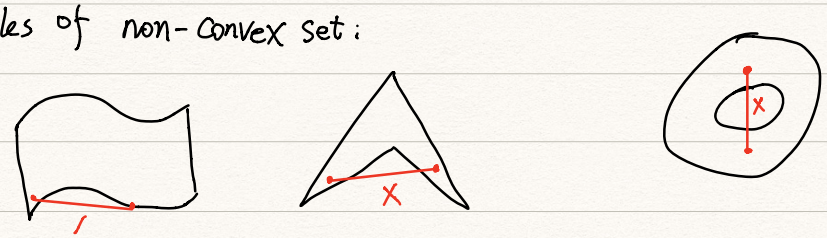
a set S is convex if

$\forall\ x, y \in S$, the line segment $\overline{xy}$ in S.

Examples of convex set:



Examples of non-convex set:



More examples of convex and non-convex sets:

① Solutions of linear equation form a convex set, i.e.,

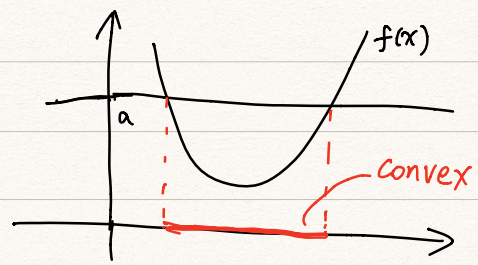
$S = \{x \in \mathbb{R}^n \mid Ax = b,\ \text{where } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^{m \times 1}\}$ is convex

for any A and b.

② Level set of a convex function is a convex set, i.e.,

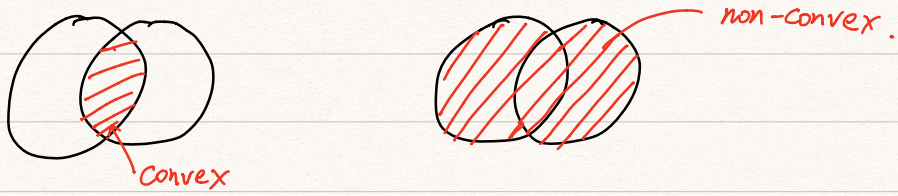
$S = \{x \in \mathbb{R}^n \mid f(x) \le a\}$, where $f: \mathbb{R}^n \to \mathbb{R}$ is convex and $a \in \mathbb{R}$, is

convex.



③ Intersection of convex sets is convex.

Union of convex sets is generally not convex.



- If $S \subset \mathbb{R}^n$ is a convex set and $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function,

  then $\min_{x \in \mathbb{R}^n} f(x)$ s.t. $x \in S$

  is a convex optimization.

- Indicator function of a convex set.

Let $S \subset \mathbb{R}^n$ be convex. We define a function $I_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

$$I_S(x) = \begin{cases} 0 & \text{if } x \in S \\ +\infty & \text{if } x \notin S \end{cases}$$

This function is called the indicator function of S.

Theorem: If S is a closed convex set, then I_S is a convex function.

proof. Let $x, y \in \mathbb{R}^n$ and $t \in [0,1]$.

Case 1: Both $x, y \in S$.

Then $tx + (1-t)y \in S$ due to the convexity of S.

So, $I_S(tx + (1-t)y) = 0$ and $I_S(x) = I_S(y) = 0$.

Therefore, $I_S(tx + (1-t)y) \leq t\, I_S(x) + (1-t)\, I_S(y)$.

Case 2: $x \notin S$ and $y \in S$.

If $t \neq 0$, $t I_S(x) + (1-t)\, I_S(y) = +\infty \geq I_S(tx + (1-t)y)$

If $t=0$, $\quad tI_S(x)+(1-t)I_S(y)=0 \geq 0=I_S(tx+(1-t)y)$

Case 3: $x\in S$, and $y\notin S$

the same as Case 2

Case 4: $x\notin S$, $y\notin S$.

$$tI_S(x)+(1-t)I_S(y)=+\infty \geq I_S(tx+(1-t)y)$$

Thus, in all cases, $I_S(tx+(1-t)y)\leq tI_S(x)+(1-t)I_S(y)$ $\boxtimes$.

- Sub-differential of indicator function

Let $S\subset\mathbb{R}^n$ be convex. Obviously, I_S is non-differentiable

- If x is in the interior of S, then I_S is differentiable and

$$\partial I_S(x)=\{0\}.$$

- If x is on the boundary of S, then I_S is non-differentiable.

$$\partial I_S(x)=\{u\in\mathbb{R}^n \mid I_S(y)\geq I_S(x)+\langle u,y-x\rangle \quad \forall y\}$$

Since for any $y\notin S$, $I_S(y)=+\infty$,

$$I_S(y)\geq I_S(x)+\langle u,y-x\rangle \quad \text{for all } u\in\mathbb{R}^n.$$

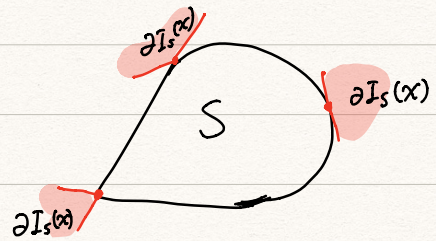
Therefore, $\quad \partial I_S(x)=\{u\in\mathbb{R}^n \mid I_S(y)\geq I_S(x)+\langle u,y-x\rangle,\ \forall y\in S\}$

$$\partial I_S(x)=\{u\in\mathbb{R}^n \mid 0\geq\langle u,y-x\rangle \quad \forall y\in S\}$$

- If x is not in S, then we define $\partial I_S(x)=$ the empty set.

In summary,

$$\partial I_S(x)=\begin{cases}\text{empty set} & \text{if } x\notin S\\ 0 & \text{if } x \text{ in the interior of } S\\ \{u\in\mathbb{R}^n \mid \langle u,y-x\rangle\leq 0,\ \forall y\in S\} & \text{if } x \text{ on boundary of } S\end{cases}$$

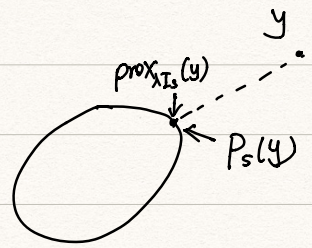


- proximity of the indicator function / projection

By definition,

$$\begin{aligned}\mathrm{prox}_{\lambda I_S}(y) &= \arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|x-y\|_2^2 + \lambda I_S(x) \\ &= \arg\min_{x\in S} \tfrac{1}{2}\|x-y\|_2^2 \\ &= \arg\min_{x\in S} \|x-y\|_2\end{aligned}$$

i.e., $\mathrm{prox}_{\lambda I_S}(y)$ is the point on S that is the closest to y, called the projection of y onto S.



From Homework 6, $\mathrm{prox}_{\lambda I_S}(y)$ always exists and is unique.

Furthermore, let $x = \mathrm{prox}_{\lambda I_S}(y)$, and Fermat's lemma gives

$$0 \in x - y + \lambda \partial I_S(x) \iff y - x \in \lambda \partial I_S(x)$$

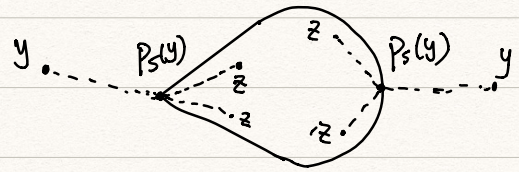
If $y \notin S$, then $y - x \in \lambda \partial I_S(x) = \partial I_S(x)$, i.e.

$$\langle y-x, z-x\rangle \le 0 \quad \forall z \in S.$$

If $y \in S$, then $x = y$ because $0 \in \lambda \partial I_S(y) \quad \forall y \in S$.

In both cases, $\langle y-x, z-x\rangle \le 0 \quad \forall z \in S$.

Theorem: The projection of y onto S, (i.e., $\mathrm{prox}_{\lambda I_S}(y)$), denoted by $P_S(y)$, always exists and is unique. Furthermore, $P_S(y)$ is the projection if and only if $\langle y - P_S(y), z - P_S(y)\rangle \le 0 \quad \forall z \in S$



Example 1: $S = \{x \mid \|x\|_2 \le 1\}$

S is closed and convex.

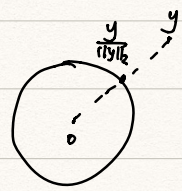
- If $\|y\|_2 \le 1$, then $y \in S$ so that $P_S(y) = y$.
- If $\|y\|_2 > 1$, then $y \notin S$, and

$$\forall z \in S, \quad \left\langle y - \frac{y}{\|y\|_2}, z \right\rangle = \left(1 - \frac{1}{\|y\|_2}\right)\langle y, z\rangle \le \left(1 - \frac{1}{\|y\|_2}\right)\|y\|_2 \|z\|_2$$

$$\le \left(1 - \frac{1}{\|y\|_2}\right)\|y\|_2 = \left(1 - \frac{1}{\|y\|_2}\right)\left\langle y, \frac{y}{\|y\|_2}\right\rangle = \left\langle y - \frac{y}{\|y\|_2}, \frac{y}{\|y\|_2}\right\rangle$$

i.e., $\left\langle y - \frac{y}{\|y\|_2}, z - \frac{y}{\|y\|_2}\right\rangle \le 0$

Therefore, $P_S(y) = \frac{y}{\|y\|_2}$.



Altogether
$$P_S(y) = \begin{cases} y & \text{if } y \in S \\ \frac{y}{\|y\|_2} & \text{if } y \notin S \end{cases}$$

Example 2: $S = \{x \in \mathbb{R}^n \mid \|x\|_\infty \le 1\}$.

S is convex and closed.

Since $S = \{x \in \mathbb{R}^n \mid |x_i| \le 1 \ \forall i = 1, 2, \cdots, n\}$, let $\Omega = \{\alpha \in \mathbb{R} \mid |\alpha| \le 1\}$

$$I_S(x) = \sum_{i=1}^{n} I_\Omega(x_i)$$

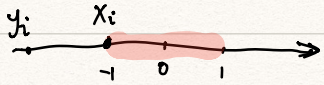
Therefore, $P_S(y) = \arg\min_{x \in S} \|y - x\|_2^2 = \arg\min_{x \in \mathbb{R}^n} \left(\|y - x\|_2^2 + I_S(x)\right)$

$$= \arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} \left((y_i - x_i)^2 + I_\Omega(x_i)\right)$$

Therefore, $[P_S(y)]_i = \arg\min_{x_i \in \mathbb{R}} (y_i - x_i)^2 + I_\Omega(x_i)$

$$= \arg\min_{|x_i| \le 1} (y_i - x_i)^2$$

$$= \begin{cases} -1 & \text{if } y_i \le -1 \\ y_i & \text{if } -1 \le y_i \le 1 \\ 1 & \text{if } y_i \ge 1 \end{cases} \equiv P_{[-1,1]}(y_i)$$



Thus, $P_S(y) = \begin{bmatrix} P_{[-1,1]}(y_1) \\ \vdots \\ P_{[-1,1]}(y_n) \end{bmatrix}$

We can check $\langle y - P_S(y), z - P_S(y)\rangle \le 0 \quad \forall z \in S$.

§ 4.3.2. Projected gradient descent

Consider $\min_{x \in S} f(x)$, where f is smooth, convex and $S \subset \mathbb{R}^n$ is convex.

This optimization is equivalent to

$$\min_{x \in \mathbb{R}^n} f(x) + I_S(x)$$

We apply Forward-Backward Splitting (FBS) to solve it:

$$x^{(k+1)} = \mathrm{prox}_{\alpha_k I_S}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right),$$

i.e.,

$$\boxed{x^{(k+1)} = P_S\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)}$$

This is known as Projected Gradient Descent. (PGD)

From the theorem in the convergence of FBS,

If $0 < \alpha_l \le \alpha_k \le \alpha_u < \frac{2}{M}$, where $M \in \mathbb{R}$ satisfies $\|\nabla^2 f(x)\|_2 \le M$,

then: ① Since $x^{(k)} \in S$, $\forall k$, $f(x^{(k)}) + I_S(x^{(k)}) = f(x^{(k)})$, $\forall k$,

$\{f(x^{(k)})\}_k$ is monotonically non-increasing.

② If $\|x^{(k)}\|_2 \le B$ $\forall k$ (This is true if, e.g., S is bounded)

then $\lim_{k \to +\infty} f(x^{(k)}) = \min_{x \in S} f(x)$.

Example 1: Non-negative least squares

$$\min_{x \ge 0} \|Ax - b\|_2^2$$

Let $S = \{x \in \mathbb{R}^n \mid x \ge 0\}$. Then PGD gives

$$x^{(k+1)} = P_S\left(x^{(k)} - \alpha_k A^T(Ax^{(k)} - b)\right),$$

where P_S is the projection onto S given by

$$P_S(y) = \begin{bmatrix} \max(y_1, 0) \\ \max(y_2, 0) \\ \vdots \\ \max(y_n, 0) \end{bmatrix}.$$