By a convex optimization (볼록최적화) we mean an optimization problem of minimizing a convex function or maximizing a concave function over a convex set. A typical form of convex optimzation is

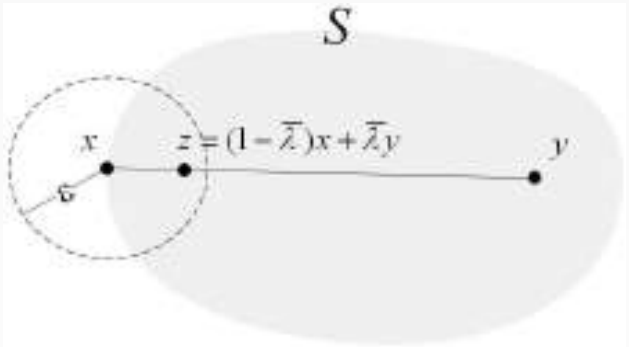
$$\begin{array}{ll} \min & \text{convex } f(x) \text{ or } \max \text{ concave } f(x) \\ \text{s.t.} & \text{convex } g_i(x) \leq 0, \text{ or} \\ & \text{concave } g_i(x) \geq 0, \ i = 1, \ldots, m, \\ & \text{affine } h_j(x) = 0, \ j = 1, \ldots, p. \end{array} \tag{8.13}$$

- The computational efforts for solving an optimization problem vary significantly depending on the characteristics of the functions in the objective or constraints. A general nonlinear program may require an astronomical scale of time and memory to obtain an optimal solution.
- A convex optimization is easy to solve, *polynomially solvable*.
  "In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity." - Rockafellar
- Prevalent! Many real problems can be formulated as a convex optimization problem such as LP, QP, SDP, etc. It is important to recognize if the given problem can be formulated or approximated by a convex optimization problem.

Theorem 8.5

Any local optimal solution of a convex optimization problem is optimal.

Proof: Let x be a local optimal solution: there is $\epsilon > 0$ such that $f(x) \leq f(z)$, $\forall z \in B_\epsilon(x) \cap S$. Assume on the contrary that there is $y \in S$: $f(x) > f(y)$. On the line segment connecting x and y, there is $z \in B_\epsilon(x)$. Let the corresponding coefficient be $\bar{\lambda}$: $z = x + \bar{\lambda}(y - x) = (1 - \bar{\lambda})x + \bar{\lambda}y$.

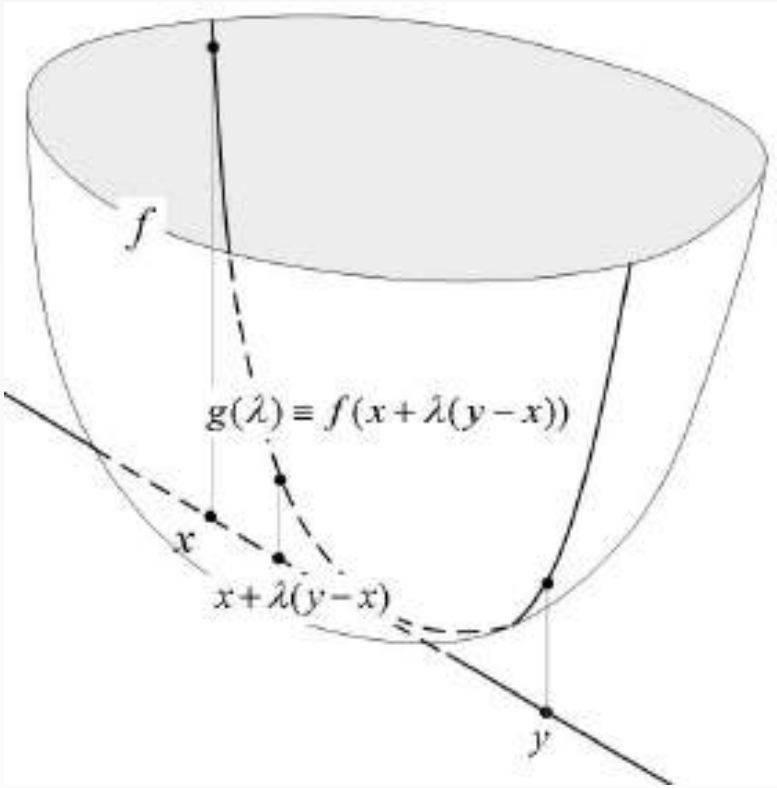


Then

$$f(z) = f((1-\bar{\lambda})x + \bar{\lambda}y) \leq (1-\bar{\lambda})f(x) + \bar{\lambda}f(y) \text{ (from convexity of } f) \\ < (1-\bar{\lambda})f(x) + \bar{\lambda}f(x) \text{ (assumption } f(y) < f(x)) = f(x). \tag{8.14}$$

A contradiction. □

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then its restriction on any line is also a (one-dimensional) convex function. Conversely, if the restriction of a function to any line is convex, the function is convex on $\mathbb{R}^n$. See the figure below.



It means to show the convexity of f, it suffices to show that for any points x, y of $\mathbb{R}^n$, $g(\lambda) = f(x + \lambda(y - x))$ is convex function in λ.

Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$.

Theorem 9.1

If f is convex, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \; \forall x, y \in \mathbb{R}^n. \tag{9.15}$$

The converse is also true.

Proof ($\Rightarrow$) Take any points x, $y \in \mathbb{R}^n$. Since f is convex, for $0 < \lambda < 1$, $f(x + \lambda(y - x)) \leq (1 - \lambda) f(x) + \lambda f(y)$, or

$$f(y) \geq f(x) + \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

If we let $g(\lambda) = f(x + \lambda(y - x))$, the last term is equal to $\frac{g(\lambda) - g(0)}{\lambda - 0}$ which converges to $g'(0)$ as $\lambda \to 0$. But $g'(0) = \nabla^T f(x)(y - x)$ and hence (9.15) follows. $\square$

Proposition 10.1

A twice-differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if $f''(x) \geq 0 \ \forall \ x$.

Proof ($\Rightarrow$) Since f is convex, the first-order condition implies for all x, y with $x < y$, $f(y) \geq f(x) + f'(x)(y - x)$, or $\frac{f(y)-f(x)}{y-x} \geq f'(x)$. Similarly, $f'(y) \geq \frac{f(y)-f(x)}{y-x}$. Hence f' is monotone increasing and thus $f''(x) \geq 0$. ($\Leftarrow$) By Taylor's theorem, for any two points $x < y$, there is $x \leq z \leq y$ such that $f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(z)(y - x)^2$. By the assumption, it implies $f(y) \geq f(x) + f'(x)(y - x)$. By the first-order condition of convexity, f is convex. $\square$

For the case $f : \mathbb{R}^n \to \mathbb{R}$, consider any x and $y \in \mathbb{R}^n$, and $g(\lambda) := f(x + \lambda(y - x))$. Then

$$g''(\lambda) = (y - x)^T \nabla^2 f(x + \lambda(y - x))(y - x) \geq 0. \tag{10.16}$$

Since x and y are arbitrary, it implies that for any $x \in \mathbb{R}^n$, we have

$$z^T \nabla^2 f(x) z \geq 0, \forall z \tag{10.17}$$

Definition 10.2

A symmetric matrix Q is said to be positive semidefinite (PSD) if $z^T Q z \geq 0$ for every z.

Proposition 10.3

A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its Hessian $\nabla^2 f(x)$ is PSD.

Exercise 10.4

- *Prove the function are convex:* $f(x) = c + d^T x$, $f(x) = c + d^T x + \frac{1}{2} x^T Q x$ *where* Q *is PSD.*
- *Negative entropy* $f(x) = x \log x$ *defined on* $\mathbb{R}_{++}$ *is convex.*
- $f(x, y) = x^4 + x^2 y^2$ *is convex on* $\left\{(x, y) \in \mathbb{R}^2 | x \leq y \leq 0\right\}$.
- *Sketch the graphs of* $f(x) = \frac{x^2}{1-|x|}$ *and* $f(x) = |x| - \ln(1 + |x|)$ *and find the range on which each* f *is convex.*
- *Exponential* e^{ax} $a \in \mathbb{R}$ *is convex.*
- x^a *defined on* $\mathbb{R}_{++}$ *is convex if either* $a \geq 1$ *or* $a \leq 0$, *and concave if* $0 \leq a \leq 1$.
- *If* $p \geq 1$, $|x|^p$ *is convex.*
- $\log x$ *is concave on* $\mathbb{R}_{++}$.

Theorem 11.1

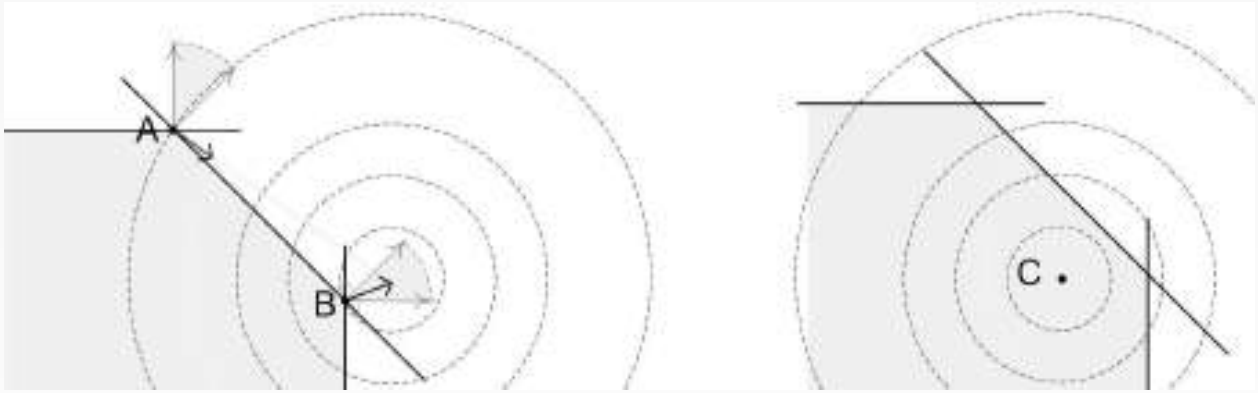
Feasible direction theorem(가능방향 정리) *Let f be a differentiable convex function defined on a convex set $F \subseteq \mathbb{R}^n$. A point $x \in F$ is a minimizer of f over F if and only if for every feasible direction y of x, $\nabla f(x)^T y \geq 0$.*

Proof: The necessity follows from Proposition 4.2 (irrespective of convexity).

For sufficiency, take any point $z \in F$ such that $z \neq x$. By the first order condition of convexity, $f(z) \geq f(x) + \nabla f(x)^T (z - x)$. Since F is convex, $y - x$ is a feasible direction, the assumption implies $\nabla f(x)^T (z - x) \geq 0$ and $f(z) \geq f(x)$. Hence, x is optimal. $\square$

Corrolary 11.2

An interior solution x of a convex optimization is optimal if and only if $\nabla f(x) = 0$.

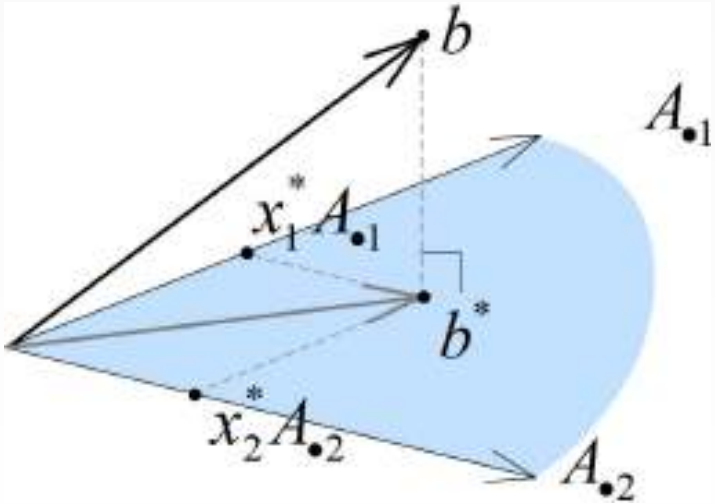


A point x satisfying $\nabla f(x) = 0$ is called a stationary point) of f.

In case a linear equality system $Ax = b$ has no solution, we may pursue a point x whose error is minimized. An error can be defined naturally by using 2-norm $\|\cdot\|_2$. Then the problem is an optimization problem.

$$\min \|Ax - b\|_2. \tag{11.18}$$

(11.18) is the problem of finding a point b^* in the column space of A which is closest to b. b^* is called the projection of b onto the row space of A. An optimal solution $x_1^*, \ldots, x_n^*$ of (11.18) is the coefficients in a linear combination of $A_{\cdot 1}, \ldots, A_{\cdot n}$ representing b^*.



Since $\|Ax-b\|_2^2 = (Ax-b)^T(Ax-b) = x^TA^TAx - 2b^TAx + b^Tb$, (11.18) can be formulated as the unconstrained problem minimizing the quadratic function $f(x) = x^TA^TAx - 2b^TAx + b^Tb$ Since $\nabla^2 f(x) = A^TA$ is PSD matrix (why?), the problem is an unconstrained convex optimization.

Optimal solutions are attained at a stationary point:

$$2A^TAx - 2A^Tb = 0. \tag{11.19}$$

Therefore optimal solutions are exactly the solutions of (11.19). If the columns of A are linearly independent, A^TA is invertible, and the solution is unique, $x^* = (A^TA)^{-1}A^Tb$,

$$b^* = A(A^TA)^{-1}A^Tb.$$

We call $A(A^TA)^{-1}A^T$ the *projection matrix onto the column space of* A.

The least-square method is used to compute regression coefficients. Suppose we have observed the values of independent variables x and dependent variables y as follows.

x	1	2	3
y	1.8	3	4.2

For a linear regression model $y(x) = ax + b$, the differences between prediction and observation are

$$\begin{aligned} y(1) - 1.8 &= a + b - 1.8 \text{ for } x = 1, \\ y(2) - 3 &= 2a + b - 3 \text{ for } x = 2, \text{ and} \\ y(3) - 4.2 &= 3a + b - 4.2 \text{ for } x = 3. \end{aligned}$$

Our problem is to find the coefficients a and b minimizing the error in 2-norm,

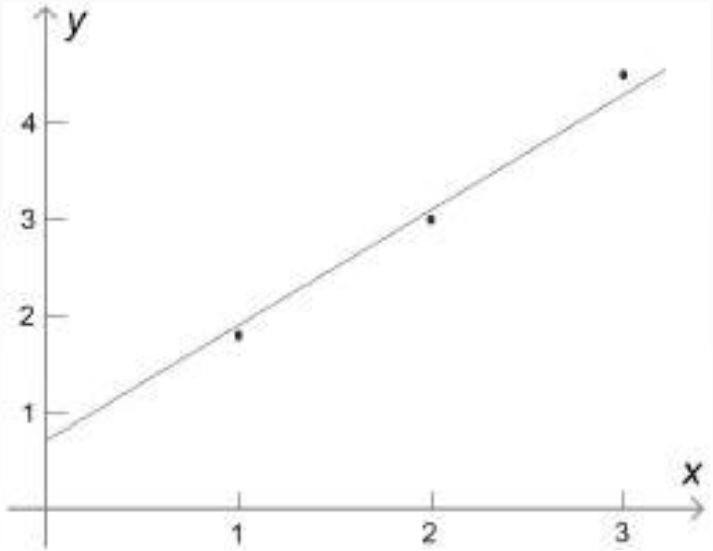
$$\min \left\| \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 1.8 \\ 3 \\ 4.2 \end{bmatrix} \right\|_2^2 .$$

Then

$$\begin{bmatrix} a \\ b \end{bmatrix} = \left(\begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1.8 \\ 3 \\ 4.2 \end{bmatrix}$$

$$= \frac{1}{6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix} \begin{bmatrix} 20.4 \\ 9 \end{bmatrix} = \begin{bmatrix} 1.2 \\ 0.7 \end{bmatrix}.$$

The obtained linear model is $y = 1.2x + 0.7$.



Exercise 11.3

Suppose we have observed the values of independent and dependent variables as follows:

x	1	2	3	4
y	3	13	20	38

We have chosen a quadratic regression model $y(x) = ax^2 + c$. *Compute a 2-norm error minimizing regression coefficients* a *and* c.

Repeat the problems using the following norms instead of 2-norm $\|\cdot\|_2$.

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}, \tag{11.20}$$

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|. \tag{11.21}$$

Proposition 12.1

A feasible solution x^* *of a convex optimization* $\min\{f(x) : g(x) \equiv (g_1(x), \ldots, g_m(x)) \geq 0\}$ *is optimal if there is* λ^* *satisfying*

$$\begin{array}{c} \nabla f(x^*) - \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0, \\ \lambda^* \geq 0, \\ \lambda_i^* = 0, \forall i \notin A(x^*) \ \Big(\Leftrightarrow \ (\lambda^*)^T g(x^*) = 0 \Big). \end{array} \qquad \text{(KKT conditions)}$$

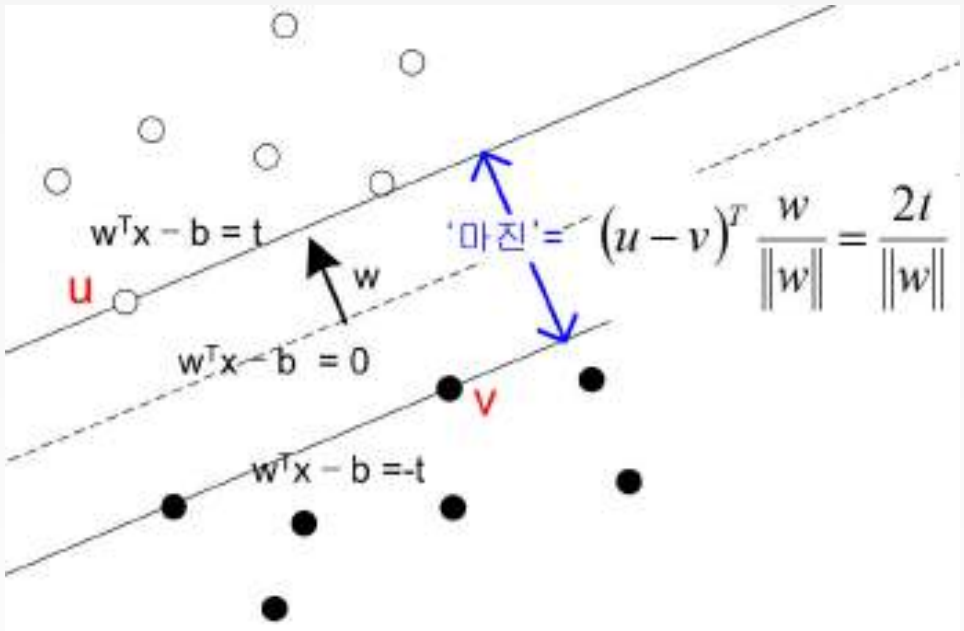
Proof: For each λ, we consider the function in x defined by $L(x;\lambda) \equiv f(x) - \lambda^T g(x)$ (called *Lagrangian* (라그랑지안)) Since g_i's are concave in x and $\lambda \geq 0$, $L(x;\lambda)$ is convex in x. Since x^* is a stationary point by the first condition, $L(x;\lambda^*)$ is minimized at $x = x^*$. In particular, for every feasible solution x,

$$L(x^*;\lambda^*) = f(x^*) - (\lambda^*)^T g(x^*) \leq f(x) - (\lambda^*)^T g(x) \leq f(x). \qquad (12.22)$$

The last inequality follows from $\lambda^* \geq 0$ and $g(x) \geq 0$. The third condition, $(\lambda^*)^T g(x^*) = 0$ implies $f(x^*) = L(x^*;\lambda^*)$, and therefore (12.22) implies $f(x^*) \leq f(x)\ \forall\ x \in F$. $\square$

Hyperplane classifier

Given two sets $\{u_i | i \in U\}$ and $\{v_i | i \in V\}$, find a hyperplane (w, b) that separates two sets by a 'maximum' margin.



Suppose the supporting hyperplanes are $w^T x = b - t$ and $w^T x = b + t$, and the support vectors are $u \in U$, $v \in V$. Then the margin is $(v - u)^T \frac{w}{\|w\|} = 2\frac{t}{\|w\|}$.

If we assume $t > 0$, then the problem of finding hyperplane with a largest margin becomes a QP:

$$\begin{array}{lll} \max & \frac{t}{\|w\|_2} & (\Leftrightarrow \min \frac{\|w\|_2}{t}) \\ \text{s.t.} & u_i^T w - b & \geq +t,\ i \in U, \\ & v_i^T w - b & \leq -t,\ i \in V, \\ \Leftrightarrow \quad \min & \|w\|_2^2 & \\ \text{s.t.} & u_i^T w - b & \geq +1,\ i \in U, \\ & v_i^T w - b & \geq -1,\ i \in V. \end{array}$$

For a general case when two sets U and V are not separable by a hyperplane, we can allow error $\xi_i \geq 0$ for each i and add a total error penalty to the objective function:

$$\begin{array}{lll} \min & \|w\|_2^2 + \gamma \sum_i \xi_i & \\ \text{s.t.} & u_i^T w - b \geq +1 - \xi_i, & i \in U, \\ & v_i^T w - b \leq -1 + \xi_i, & i \in V, \\ & \xi_i \geq 0, & i \in U, V. \end{array}$$