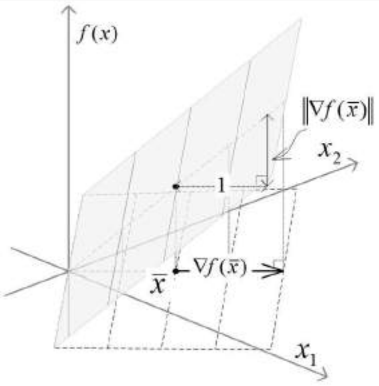
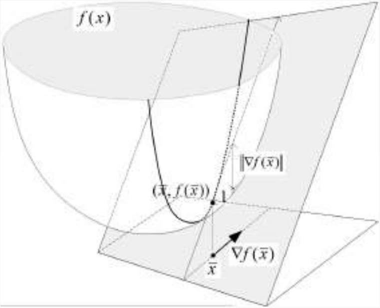
Definition 2.1

The *gradient*(기울기 벡터) of a function f at $x = \bar{x}$ is defined as

$$\nabla f(\bar{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\bar{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\bar{x}) \end{bmatrix}. \tag{2.4}$$

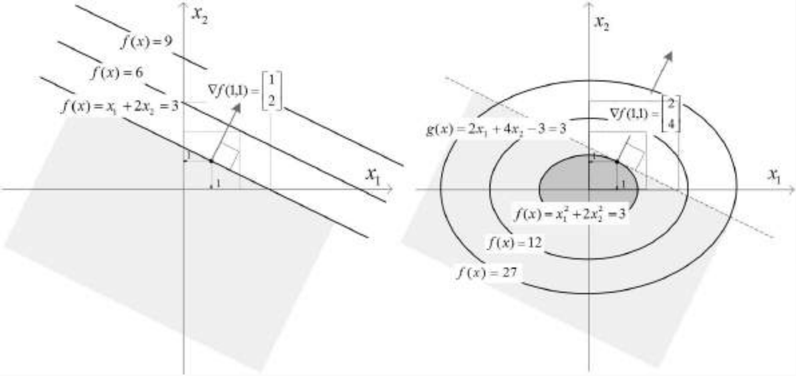


The gradient $\nabla f(x) = [1, 1]^T$ of the linear functional $f(x_1, x_2) = x_1 + x_2$ is the direction into which f increases fastest. It is normal to the contour (or level set) containing $\bar{x}$. The rate of increase is given by $\|\nabla f(1, 1)\|_2 = \sqrt{2}$.



In general, the gradient $\nabla f(\bar{x})$ of a real-valued function $f(x)$ at $x = \bar{x}$ is the same as the gradient of the linear function whose graph is the plane tangent to the graph of $f(x)$ at $(x, f(x))$. The instantaneous rate of increase of the function at $x = \bar{x}$ is largest in the direction of $\nabla f(\bar{x})$ and is equal to $\|\nabla f(\bar{x})\|_2$.

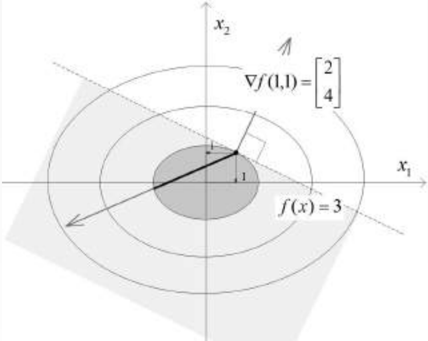
Level sets of $f(x) = x_1 + 2x_2$ and $f(x) = x_1^2 + 2x_2^2$.



Suppose the inner product y and $f(\bar{x})$ is negative. If f is linear, it decreases at a constant rate along the line from $\bar{x}$ into the direction y. A nonlinear function does not necessarily decreases at a constant rate. But, there is an open interval immediately after $\bar{x}$ along the line on which the function value is smaller than $f(\bar{x})$

Definition 3.1

A vector y is said to be a descent direction from $\bar{x}$ if $\exists\ \bar{\lambda} > 0$: $f(\bar{x} + \lambda y) < f(\bar{x})\ \forall\ 0 < \lambda < \bar{\lambda}$.



In the figure, we can see every direction from $\bar{x}$ having a negative inner product with $\nabla f(\bar{x})$ is a descent one.

- The derivative $Df(\bar{x})$ of a function $f : \mathbb{R}^3 \to \mathbb{R}^2$, $x = [x_1, x_2, x_3]^T \mapsto f(x) = [f_1(x), f_2(x)]^T$ at $x = \bar{x}$ is defined as

$$Df(\bar{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \frac{\partial f_1}{\partial x_2}(\bar{x}) & \frac{\partial f_1}{\partial x_3}(\bar{x}) \\ \frac{\partial f_2}{\partial x_1}(\bar{x}) & \frac{\partial f_2}{\partial x_2}(\bar{x}) & \frac{\partial f_2}{\partial x_3}(\bar{x}) \end{bmatrix}, \tag{4.5}$$

  a linear transformation $\mathbb{R}^3 \to \mathbb{R}^2$.
- The derivative of linear functional $c^T x$ is c^T. The derivative of a general linear function $f(x) = Ax$ is A.
- In general $Df(\bar{x})$ is the linear approximation of f around $x = \bar{x}$ whose error decreases faster than the distance from $\bar{x}$: $\|f(\bar{x} + y) - f(\bar{x}) - Df(\bar{x})y\|_2 = o(\|y\|_2)$.
- **Hessian** (헤시안) The derivative $D(\nabla f)(\bar{x})$ of the function $x \mapsto \nabla f(x)$ at $x = \bar{x}$ is called the Hessian of f at $x = \bar{x}$.

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\bar{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\bar{x}) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\bar{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\bar{x}) \end{bmatrix} \tag{4.6}$$

Proposition 4.1

If $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ are diffrentiable, their composition $f := g \circ h : \mathbb{R}^n \to \mathbb{R}^p$, $x \mapsto g(h(x))$ is also differentiable and

$$Df(x) = D(g \circ h)(x) = Dg(h(x))Dh(x).$$

- $f : \mathbb{R}^3 \to \mathbb{R}$, $x = [x_1, x_2, x_3]^T$, $Df(x) = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}] \in \mathbb{R}^{1\times 3}$.
- $g : \mathbb{R} \to \mathbb{R}^3$, $t \mapsto [g_1(t), g_2(t), g_3(t)]^T$, $Dg(t) = [g_1'(t), g_2'(t), g_3'(t)]^T \in \mathbb{R}^{3\times 1}$.
- $h := f \circ g$, $t \mapsto (f \circ g)(t) = f(g_1(t), g_2(t), g_3(t))$. Then

$$\begin{aligned} h'(t) &= Df(g(t))Dg(t) \\ &= [\tfrac{\partial f}{\partial x_1}(g(t)), \tfrac{\partial f}{\partial x_2}(g(t)), \tfrac{\partial f}{\partial x_3}(g(t))] \begin{bmatrix} g_1'(t) \\ g_2'(t) \\ g_3'(t) \end{bmatrix} \\ &= \nabla^T f(g(t))Dg(t). \end{aligned} \tag{4.7}$$

- In the case, $g(t) = x + ty$ (x, $y \in \mathbb{R}^3$), $Dg(t) = y$ and $h'(t) = \nabla^T f(x + ty)y$. We call $h'(0) = \nabla^T f(x)y$ is the *directional derivative* of f at x into y.

Chain rule extends to any finite number of functions:

$$D(f \circ g \circ h)(x) = Df(g(h(x)))Dg(h(x))Dh(x).$$

Since $h'(t) = \nabla^T f(x+ty)y$ is the composition of the three maps

$$t \mapsto \underbrace{x+ty}_{z} \mapsto \underbrace{\nabla f(\underbrace{x+ty}_{z})}_{w} \mapsto y^T \underbrace{\nabla f(x+ty)}_{w},$$

the chain rule implies

$$h''(t) = y^T \nabla^2 f(x+ty)y.$$

Proposition 4.2

Every y such that $\nabla f(\bar{x})^T y < 0$ is a descent direction.

Proof: We take it for granted for a function in a single variable. For a function f in $x \in \mathbb{R}^n$, we consider $g(\lambda) := f(\bar{x} + \lambda y)$, a function in $\lambda \in \mathbb{R}$. Then by the chain rule,

$$g'(0) = \nabla f(\bar{x})^T y > 0. \tag{4.8}$$

By the single-variable case, there is $\bar{\lambda} > 0$ such that

$$\forall\, 0 < \lambda \le \bar{\lambda},\; f(\bar{x} + \lambda y) > f(\bar{x}). \square$$

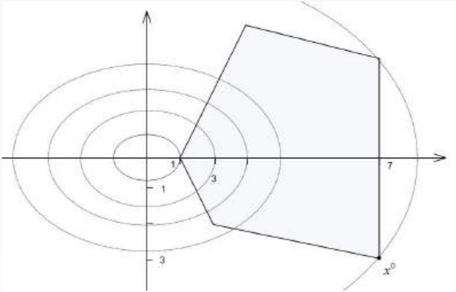
Exercise 4.3

(1) Define an ascent direction. Restate the proposition in ascent direction.
(2) Sketch the ascent directions of $f(x) = (x_1 - 2x_2)^2$ at $x = (1, 1)$.

Exercise 4.4

Compute the descent directions of the objective function from x^0.

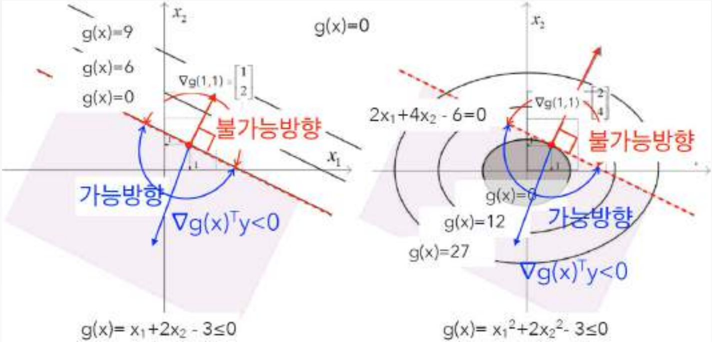
$$
\begin{array}{rllll}
\max & \frac{3}{4}x_1^2 & +x_2^2 & & \\
\text{sub.to} & 2x_1 & -x_2 & \geq 2, \\
& 2x_1 & +x_2 & \geq 2, \\
& x_1 & +4x_2 & \leq 19, \\
& x_1 & & \leq 7, \\
& x_1 & +5x_2 & \geq -8.
\end{array}
$$



Definition 5.1

If we can move from $x \in F$ into the direction y for a positive distance maintaining feasibility, i.e. $\exists\ \bar{\lambda} > 0$ such that $x + \lambda y \in F$, $\forall\ 0 < \lambda < \bar{\lambda}$, y is called a *feasible direction* of x.

If $\bar{x}$ satisfies a constraint $g(x) \leq 0$, where g is differentiable, with equality, any y such that $\nabla^T g(\bar{x}) y < 0$ is a feasible direction of $\bar{x}$.



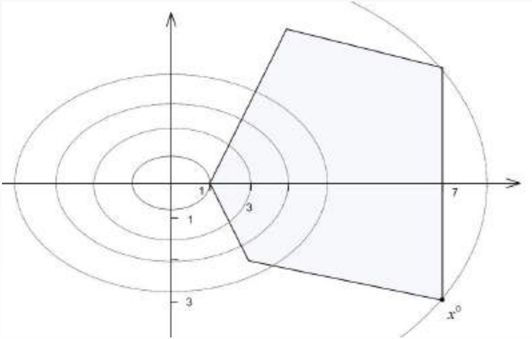
Exercise 5.2

Repeat for the constraint $g(x) \geq 0$.

If there are more than one constraints $g_i(x) \leq 0$, a direction y satisfying $\nabla g_i^T(\bar{x})y < 0$, for all i, is a feasible direction of $\bar{x}$.

Exercise 5.3

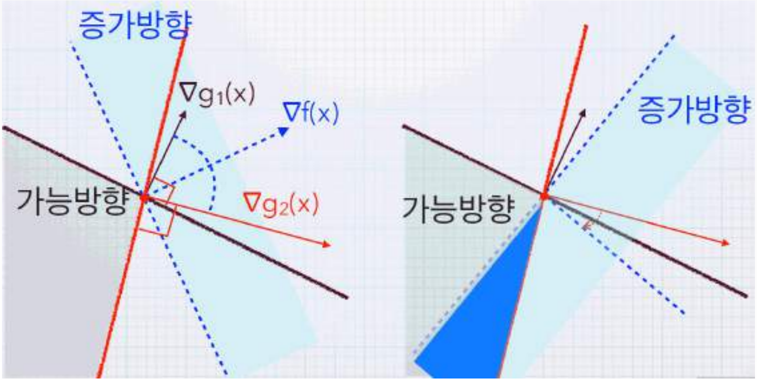
Compute the feasible directions of x^0 in the optimization problem in Exercise 4.4. Is x^0 optimal? Explain.



Definition 6.1

For min problems, Improving directions = Descent directions ∩ Feasible directions. For max problem, ...

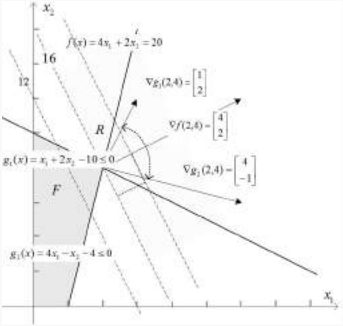
Suppose our problem is $\max \{f(x) : g_1(x) \leq 0,\ g_2(x) \leq 0\}$.



A necessary condition of optimality: Any (local) optimal solution should not have an improving direction.

Example 6.2

$$
\begin{array}{rrrrrr}
\max & f(x) = & 4x_1 & +2x_2 & & \\
\text{sub. to} & g_1(x) = & x_1 & +2x_2 & -10 & \leq 0 \\
& g_2(x) = & 4x_1 & -x_2 & -4 & \leq 0 \\
& g_3(x) = & -x_1 & & & \leq 0 \\
& g_4(x) = & & -x_2 & & \leq 0
\end{array} \tag{6.9}
$$



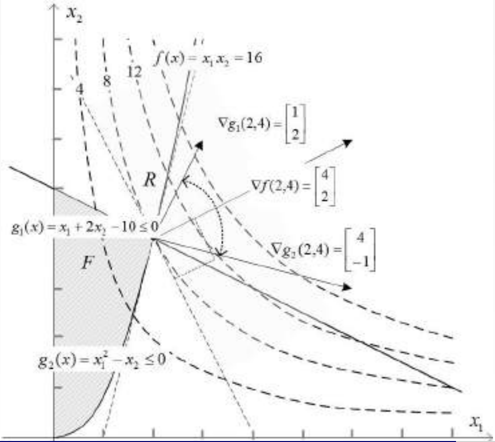
For the point $(2,4)$, any d: $[4,2]^T d > 0$ is an ascent direction. Also any d having a negative inner product with the gradients $[1,2]^T$, $[4,-1]^T$ of active constraints is a feasible direction. If $\bar{x}$ is a local optima, the two set of directions have no intersection.

If $g_1(\bar{x})$, $\nabla g_2^T(\bar{x})$ are linear indep. $\nabla g_1^T(\bar{x})y < 0$, $\nabla g_2^T(\bar{x})y < 0$ is nonempty. Then $0 \geq \sup \{\nabla f^T(\bar{x})y : \nabla g_1^T(\bar{x})y < 0, \nabla g_2^T(\bar{x})y < 0\} \Leftrightarrow 0 \geq \max \{\nabla f^T(\bar{x})y : \nabla g_1(\bar{x})y \leq 0, \nabla g_2(\bar{x})y \leq 0\}$.

By strong duality, $\Leftrightarrow \exists\, y \geq 0 : \nabla f(\bar{x}) = \nabla g_1(\bar{x})y_1 + \cdots \nabla g_m(\bar{x})y_m$, where y_i's of the inactive constrains are all 0.

The same principle applies to any nonlinear program.

$$\begin{array}{rrrll} \max & f(x) = & & x_1 x_2 & \\ \text{sub. to} & g_1(x) = & x_1 & +2x_2 - 10 & \leq 0 \\ & g_2(x) = & x_1^2 & -x_2 & \leq 0 \\ & & & x \geq 0 & \end{array} \tag{6.10}$$



Repeat the arguments for $\bar{x} = (2,4)$ to see that the necessary condition of a nonlinear program is exactly the necessary condition of the linear program obtained by the linear approximation of the problem.

Explain why either $(0,5)$ or $(1,1)$ can not be an optimal solution?

If a constraint $g_i(x) \leq 0$, $1 \leq i \leq m$ is satisfied by equality $g_i(\bar{x}) = 0$ for a feasible $\bar{x}$, it is called an active constraint of $\bar{x}$. We will denote the indices of active constraints by $A(\bar{x})$.

Proposition 7.1

Suppose $\bar{x}$ *is a local optimum of* $\max\{f(x)|g(x) \leq 0\}$. *If* $\{\nabla g_i(\bar{x}) : i \in A(\bar{x})\}$ *are linearly independent, then there is* $\lambda \in \mathbb{R}^m$ *such that*

$$\begin{gathered} \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0, \\ \lambda \geq 0, \\ \lambda_i = 0, \forall i \notin A(\bar{x}). \end{gathered} \tag{7.11}$$

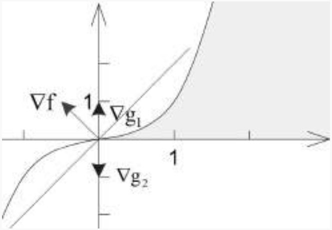
Exercise 7.2

Restate the necessary optimality condition for
$\min\{f(x)|g_i(x) \geq 0,\ 1 \leq i \leq m\}$.

Remark 7.3

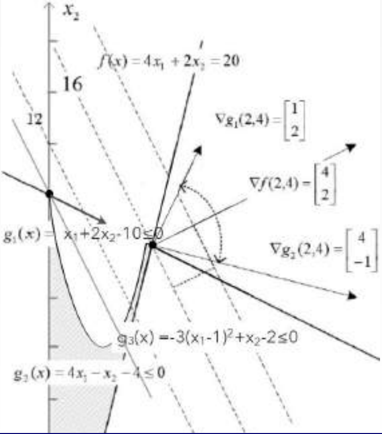
- The following example shows that the 'regularity condition' is essential. (Without it, there may be no $y : \nabla^T g_i(\bar{x})y < 0$.

$$\begin{array}{rl} \max & -x_1 + 2x_2 \\ \text{s.t.} & -x_1^3 + x_2 \le 0 \\ & -x_2 \le 0. \end{array}$$



- In the case of convex optimization, the regularity can be replaced by that " there is interior feasible solution x: $g_i(x) < 0$ for all i," Slater condition.

$$\begin{array}{llcclr} \max & f(x) = & & x_1x_2 & & \\ \text{sub. to} & g_1(x) = & x_1 & +2x_2 - 10 & \le 0 & \\ & g_2(x) = & x_1^2 & -x_2 & \le 0 & \quad (7.12) \\ & g_3(x) = & -3(x_1-1)^2 & +x_2 - 2 & \le 0 & \\ & & & x \ge 0 & & \end{array}$$



The feasible $(0, 5)$ is a local optimum but not an (global) optimum.

Definition 8.1

For reals α, the following set is called α-*sublevel* set of f:

$$C_\alpha = \{x \in \mathrm{dom} f | f(x) \leq \alpha\}.$$

Proposition 8.2

An sublevel set of a convex function is also convex. But the converse is not true.

Definition 8.3

(1) The graph of $f : \mathbb{R}^n \to \mathbb{R}$ is the set $\{(x, f(x)) | x \in \mathrm{dom} f\}$.
(2) The epigraph of $f : \mathbb{R}^n \to \mathbb{R}$ is the set $\mathrm{epi} f = \{(x, t) | x \in \mathrm{dom} f, f(x) \leq t\}$.
(3) The hypograph of $f : \mathbb{R}^n \to \mathbb{R}$ is the set $\mathrm{hyp} f = \{(x, t) | x \in \mathrm{dom} f,\ f(x) \geq t\}$.

Remark 8.4

A function is convex (concave) if and only if its epigraph (hypograph) is convex.

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

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

- 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.