Example 13.2

$$\min \quad \tfrac{1}{2}x^T P x + q^T x + r, \tag{13.25}$$

where P is a PSD matrix, $q \in \mathbb{R}^n$, and $r \in \mathbb{R}$.

- Any x^* satisfying $Px^* = -q$ is an optimal solution.
- If P is invertible, $x^* = -P^{-1}q$ is a unique optimal solution.
- If $Px = -q$ does not have a solution, (13.25) is unbounded below.

Example 13.3

$$\min \quad \|Ax - b\|_2^2 = x^T(A^TA)x - 2(A^Tb)^Tx + b^Tb. \tag{13.26}$$

The optimality conditions $A^TAx^* = A^Tb$ are called the normal equations of the least-square problem.

In iterative algorithms, we generate a minimizing sequence $x^{(k)}$, $k = 1, 2, \ldots$

$$x^{(k+1)} = x^{(k)} + \sigma^{(k)} d^{(k)},\ \sigma^{(k)} > 0,$$

where, $d^{(k)}$ is called *search direction* at iteration k, and $\sigma^{(k)}$ *step size* at iteration k.

In descent method, sequence $x^{(k)}$, $k = 1, 2, \ldots$ satisfies

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

Proposition 14.1

If f is convex, a method is descent if and only $\nabla f(x^{(k)})^T d^{(k)} < 0$.

A natural choice is then $d^{(k)} = -\nabla f(x^{(k)})$.

- Compute an initial point $x^{(0)}$.
- Until a stopping criterion is satisfied, generate x^k $k = 1, 2, \ldots$:

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

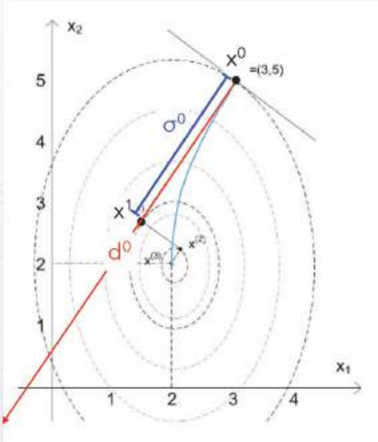
where, $\sigma^{(k)} > 0$ is called the *step size* at iteration k.

- $\sigma^{(k)} = \sigma > 0$ fixed.
- $\sigma^{(k)} = \arg\min_{\sigma > 0} f\left(x^{(k)} - \sigma \nabla f(x^{(k)})\right)$. Not practical!
- $\sigma^{(k)} = \frac{\sigma}{\sqrt{k+1}}$, for a constant $\sigma > 0$.
- In exact line search e.g. Goldstein-Armijo's rule.

Example 14.2

$$\min f(x) = 2(x_1 - 2)^2 + (x_2 - 2)^2$$

- Initial point: $(3, 5)$
- Main step: $x^{(k+1)} \leftarrow x^{(k)} + \sigma^{(k)} d^{(k)}$, where $d^k = -\nabla f(x^k)$, $\sigma^{(k)} = \text{argmin}_{\sigma > 0}\, f(x^k + \sigma d^k)$.



If f is convex around an optima solution to which initial solution is close enough, we can show, $r^{(k+1)} \leq q^k r^{(k+1)}$ for $0 < q < 1$, where $r^{(k)}$ is the distance between $x^{(k)}$ and optimal solution.

Originally proposed for finding a root of function $\phi : \mathbb{R} \to \mathbb{R}$: $\phi(t) = 0$. Relying on the linear approximation, $\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t + o(\Delta t)$, a single step is given by

$$t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)}.$$

Naturally extends to nonlinear equations: $F(x) = 0$ with $F : \mathbb{R}^n \to \mathbb{R}^n$. From linear approximation $F(x) + DF(x)\Delta x = 0$, called *Newton system*, we get an iteration

$$x^{(k+1)} = x^{(k)} - \left(DF(x^{(k)})\right)^{-1} F(x^{(k)}. \tag{14.27}$$

Adopting (14.27) to find a stationary point $\nabla f(x) = 0$ of $f : \mathbb{R}^n \to \mathbb{R}$, we get

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k). \tag{14.28}$$

The step (14.28) can be recaptured as a stationary point of the quadratic approximation of f at x^k:

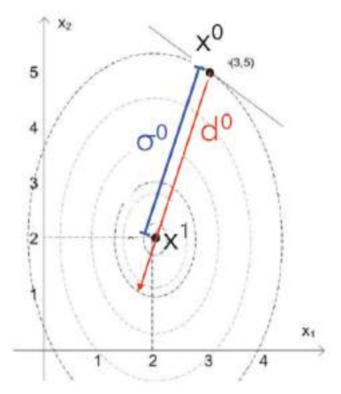
$$f(x^k) + \nabla f(x^k)(x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k). \tag{14.29}$$

Therefore, Newton method is based on quadratic approximation of f.

$$f(x) = 2(x_1-2)^2 + (x_2-2)^2 = 2x_1^2 + x_2^2 - 8x_1 - 4x_2 + 12 =$$
$$\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -8 & -4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 12 = \frac{1}{2}x^T Q x + b^T x + c$$

- $\nabla f(x) = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} -8 \\ -4 \end{bmatrix} = 0$
- $x = -Q^{-1}b = [2,2]^T$
- Main step: $x^{k+1} \leftarrow x^k + d^k$, $d^k = -\nabla^2 f(x^{(k)})^{-1}\nabla f(x)$.

Thus if f is convex quadratic function, Newton method finds optimum in a single iteration.



Advantage of Newton method.

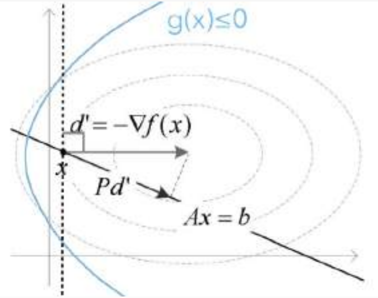
- If f is strictly convex around an optimal solution to which initial solution is close enough, it converges to an optimal solution very fast. Quadratic convergence: $r_{k+1} = cr_k^2$.

Disadvantage of Newton method.

- Breaks down in the neighborhood of x if $\nabla^2 f(x)$ is singular.
- It may not converge at all without a local convexity.

등호제약을 위한 투영 원리
등호제약식이 볼록 해집합을 가지려면 선형이 되어야 한다.

- $g(x) = Ax - b = 0$
- A의 행이 모두 선형독립이면, 벡터를 $Ax = 0$의 공간에 투영한 벡터는 다음의 관계로 주어진다: $P = I - A^T(AA^T)^{-1}A$. 최적화 문제 $\min\{||d - d'||^2 : Ad = 0\}$의 KKT 조건을 고려할 것.



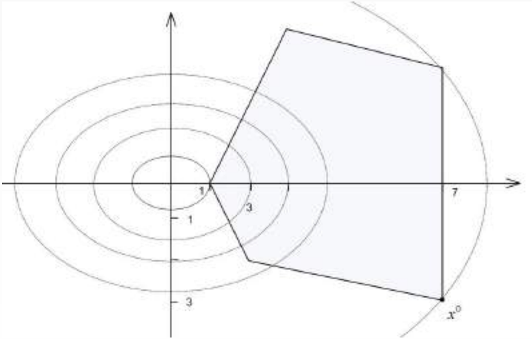
이러한 투영 방향이 원리는 다양한 방법에 적용 될 수 있다. (Projected gradient, reduced gradient, ...)

Exercise 14.3

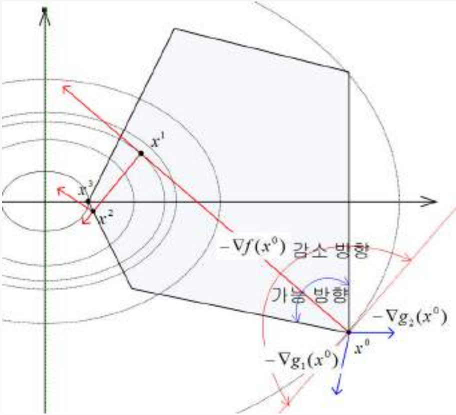
(1) $P^2 = P$임을 보여라. (2) $d = -P\nabla f(x)$ $(d \neq 0)$가 가능 하강 방향임을 보여라. (3) 위 최적화 문제의 KKT 조건을 풀어 최적해를 구하여라.

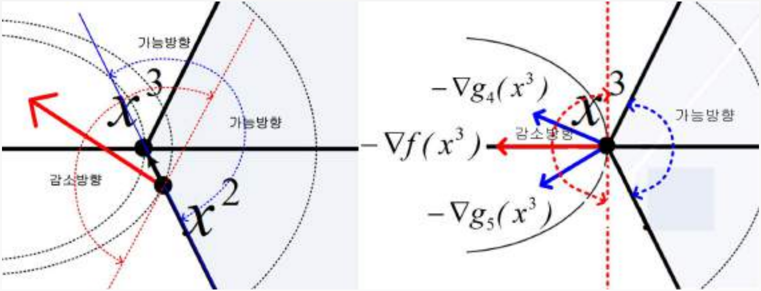
부등호 제약을 위한 개선방향 알고리듬

- 초기해 x^0 를 구한다. $k \leftarrow 0$.
- 반복단계:
  - 개선방향 d^k를 하나 구한다. 없으면 종료 현재 해 x^k 최적.
  - 이동 거리 σ^k를 정한다. $x^{k+1} \leftarrow x^k + \sigma^k d^k$.
  - $k \leftarrow k + 1$으로 놓고 반복한다.



최대 하강 (steepest descent) + 선 탐색 (line search)





Exercise 14.4

x^2에서 active constraints로 정의된 초평면 위에 음의 기울기 벡터를 투영한 방향을 구하라.

The barrier method converts an inequality constrained problem (P): $\min\{f(x) : g(x) \geq 0\}$ into an unconstrained problem by using a 'barrier' function. We assume $g(x) \geq 00$ has an interior feasible solution.

Let $I : \mathbb{R} \to \mathbb{R}$ be an indicator function: $I(u) = 0$ if $u \geq 0$, $= +\infty$ $u < 0$. Then the problem (P) is equivalent to the unconstrained problem $\min\ f(x) + \sum_i I(g_i(x))$.

But the reformulation is not useful as indicator function I is non-differentiable.

The idea is to approximate I with a smooth function $B(\cdot)$ called a *barrier function*.

Example 15.2

For a single constraint $g(x) \geq 0$, $B(h(x)) = -\frac{1}{t}\ln(g(x))$, for a parameter $t > 0$, is called *log barrier*. For instance, $-\frac{1}{t}\ln x$, $t > 0$, is the log barrier of the constraint $x \geq 0$ of single variable x. Consider a simple single variable optimization problem, (P) $\min\{f(x) : x \geq 0\} \approx$ and corresponding barrier problem (BP) $\min\{f(x) - \frac{1}{t}\ln x : x \in \mathbb{R}\}$. We can show optimal solutions of (BP), $x^*(t) \to x^*$ as $t \to +\infty$. Intuitively, this is because the weight of the barrier term gets smaller and the optimality of $f(x)$ is more stressed as t grows.

Example 15.3

Let $g_1(x) = x \geq 0$, $g_2(x) = -x + 10 \geq 0$. Then the log barrier $B(g_1(x)) + B(g_2(x) = -\frac{1}{t}(\ln x + \ln(10 - x))$ becomes closer to the indicator function $I(g_1(x)) + I(g_2(x))$ as $t \to \infty$.