1 Affine and Convex Sets

1.1 Relative interior

Affine dimension of C is $\dim \mathbf{aff}\, C$.
$\mathbf{relint}\, C = \{x \in C \mid B(x,r) \cap \mathbf{aff}\, C \subseteq C \text{ for some } r > 0\}$
Think of relative interior as the analogous version of normal interior, just as the universe is compressed from the original ambient set to the affine hull. Assume Euclidean norm and Euclidean ball for understanding.

1.2 Cones

A set C is a *cone* if $\forall x \in C$ and $\theta > 0$ we have $\theta x \in C$. C is convex cone if it satisfies both definitions.
Conic combination is a point of the form $\sum_{i=1}^{k} \theta_i x_i$, where $\theta_i \geq 0$.
Prove by induction that C is a **convex** cone $\iff$ C contains all possible conic combinations.

1.3 Operations preserving convexity

Intersections of convex sets: **Every** closed convex set is the intersection of all halfspaces containing it.
Affine function: function $f : \mathbf{R}^n \to \mathbf{R}^m$ with the form $f(x) = Ax + b$, where $A \in \mathbf{R}^{m \times n}$ and $B \in \mathbf{R}^m$
If $S \in \mathbf{R}^n$ is convex, then $f(S)$ is convex. If $P \in \mathbf{R}^m$ is convex, then $f^{-1}(P)$ is convex. Examples include scaling, translation, projection, sum, partial sum.

1.4 Generalised order

Proper cone: a cone K which is convex, closed, solid(interior not empty), pointed(if $x \in K$ and $-x \in K$, then $x = 0$), then we can define $x \preceq_K y \Leftrightarrow y - x \in K$, and $x \prec_K y \Leftrightarrow y - x \in \mathbf{int} K$.
Some properties: preservation under addition, nonnegative scaling and limits, transitivity, reflexivity, anti-symmetric.

1.5 Separating hyperplane theorem

Theorem: suppose C and D are convex, nonempty and disjoint sets, $\exists a \neq 0$ and b such that$a^T x \leq b\ \forall x \in C$ and $a^T x >= b\ \forall x \in D$. If at least one of C and D is bounded, and both are closed, the *strict* inequalities hold. *Intuition*: Construct singleton/finite sets to satisfy strict separation condition.
Proof: Fix two points in C and D that achieves the distance of two sets, then $a = d - c$, $b = \frac{1}{2}(|d|_2^2 - |c|_2^2)$. Prove by contradiction that it is true.

1.6 Supporting hyperplane

Suppose $C \in \mathbf{R}^n$, and $x_0 \in \mathbf{bd} C = \mathbf{cl} C \setminus \mathbf{int} C$, if $\exists a \neq 0$ such that $a^T x \leq a^T x_0\ \forall x \in C$, then the hyperplane $a^T x = a^T x_0$ is a *supporting hyperplane* to C at point x_0. For every nonempty convex set C and any $x_0 \in \mathbf{bd} C$, such a plane exists.

2 Convex functions

2.1 First-order conditions

If f is differentiable, then f is convex iff $\mathbf{dom} f$ is convex and $f(y) \geq f(x) + \nabla f(x)^T (y - x)\ \forall x, y \in \mathbf{dom} f$.
Note that zero gradient in this case implies a global minimum, and *the* global minimum if f is strictly convex. Strict convexity can be implied by strict inequality.

2.2 Second-order conditions

Hessian matrix: $H_{ij} = \frac{\delta f}{\delta x_i \delta x_j}$. f is convex iff $\mathbf{dom} f$ is convex and its Hessian is positive semidefinite.

2.3 Examples

Exponential: e^{ax} is convex on $\mathbf{R} \forall a \in \mathbf{R}$.
Powers: x^a is convex on $\mathbf{R}^{++}$ when $a \geq 1 \text{or} a \leq 0$, and concave for $a \in [0, 1]$.
Powers of absolute value: $|x|^p$, for $p \geq 1$, is convex on $\mathbf{R}$.
Logarithm: $\log x$ is concave on $\mathbf{R}^{++}$.
Negative entropy: $x \log x$ is strictly convex on $\mathbf{R}^{++}$.
Other examples include norms, max, quadratic-over-linear, log-sum-exp, geometric mean, log-determinant.

2.4 Sublevel sets

The α-sublevel set of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as $C_\alpha = \{x \in \mathbf{dom} f : f(x) \leq \alpha\}$. It is convex if f is convex for any value of α. Analogous α-superlevel set by $\{x \in \mathbf{dom} f : f(x) \geq \alpha\}$ is convex if f is concave. Converse is not true. Consider $-e^x$.

2.5 Epigraph

$\mathbf{graph}(f) = \{(x, f(x)) : x \in \mathbf{dom} f\}$
$\mathbf{epi}(f) = \{(x, t) : x \in \mathbf{dom} f, t \geq f(x)\}$
$\mathbf{hypo}(f) = \{(x, t) : x \in \mathbf{dom} f, t \leq f(x)\}$
f is convex iff $\mathbf{epi}(f)$ is convex, and is concave iff $\mathbf{hypo}(f)$ is convex.

2.6 Operations preserving convexity

Conic combinations of convex functions
Composition with affine functions
Pointwise max/sup: geometrically, the epigraph of the max/sup function is the intersection of epigraphs of all component functions, which are all convex. **Intuition**: establish convexity of a function by construction from max/sup of other more obvious convex functions.
Composition: Assume **twice differentiability**, check cases to ensure the second derivative is non-negative. **Some compositions**: If g is convex, then $\exp g(x)$ is convex. If g is concave and positive, then $\log g(x)$ is concave. If g is concave and positive, then $\frac{1}{g(x)}$ is convex. If g is convex and non-negative and $p \geq 1$, then $g(x)^p$ is convex. If g is convex then $-\log(-g(x))$ is convex on $\{x \in \mathbf{dom} g : g(x) < 0\}$.
Minimization: if f is convex in (x, y), and C is convex and nonempty, then $g(x) = \inf_{y \in C} f(x, y)$ is convex. $\mathbf{dom} g = \{x : \exists y \in C, (x, y) \in \mathbf{dom} f\}$

3 Conjugate functions

3.1 Definitions

Let f: $\mathbf{R}^n \to \mathbf{R}$, then *conjugate* of f, which is f^*: $\mathbf{R}^n \to \mathbf{R}$ is defined as $f^*(y) = \sup_{x \in \mathbf{dom} f}(y^T x - f(x))$. The domain of f^* consists of $y \in \mathbf{R}^n$ where the supremum is finite.
f^* is convex in y as it is the pointwise supremum of a family of convex(indeed affine) functions of y.

3.2 Examples

Affine functions: $f(x) = ax + b$, $x \in \mathbf{R}$, then $f^*(y) = \sup_{x \in \mathbf{R}}\{xy - ax - b\} = \sup_{x \in \mathbf{R}}\{x(y-a)\} - b$, hence $f^*(y) = -b$ when $y = a$ and $+\infty$ otherwise.
Negative logarithm: $f(x) = -\log(x), x \in \mathbf{R}^{++}$. $f^*(y) = \sup_{x \in \mathbf{R}^{++}}\{xy + \log(x)\}$, hence $f^*(y) = +\infty$ when $y \geq 0$ and $f^*(y) = -\log(-y) - 1$ when $y < 0$ by differentiation. **Intuition**: divide cases for y, draw graphs to see the sum and find potential extremum by calculus.
Exponential: $f(x) = e^x$, $x \in \mathbf{R}$, then $f^*(y) = \sup_{x \in \mathbf{R}}\{xy - e^x\}$, hence $f^*(y) = +\infty$ when $y < 0$, and $f^*(y) = y\log(y) - y$ when $y > 0$ and $f^*(0) = 0$.
Inverse: $f(x) = \frac{1}{x}$, $x \in \mathbf{R}^{++}$, then $f^*(y) = \sup_{x \in \mathbf{R}^{++}}\{xy - \frac{1}{x}\}$, hence $f^*(y) = +\infty$ when $y > 0$, $f^*(y) = -2(-y)^{\frac{1}{2}}$ when $y < 0$, and $f^*(y) = 0$ when $y = 0$.
Log-sum-exp function: $f(x) = \log(\Sigma_{i=1}^n e^{x_i})$, $x \in \mathbf{R}^n$, then $\mathrm{dom} f^* = \{y \in \mathbf{R}^n : y_i \geq 0, \Sigma_{i=1}^n y_i = 1\}$, and $f^*(y) = \Sigma_{i=1}^n y_i \log(y_i)$

3.3 Properties

By definition, we have $f^*(y) \geq x^T y - f(x) \to f^*(y) + f(x) \geq x^T y\ \forall x \in \mathbf{dom} f$, which is Fenchel's inequality.
Scaling and composition with affine function: For $a > 0$ and $b \in \mathbf{R}$, the conjugate of $g(x) = af(x) + b$ is $g^*(y) = af^*(\frac{y}{a}) - b$, and $\mathbf{dom} g^* = \{y \in \mathbf{R}^n : \frac{y}{a} \in \mathbf{dom} f^*\}$.
Sum of independent functions: $f(u, v) = f_1(u) + f_2(v)$, then $f^* = f_1^* + f_2^*$.
Conjugate of conjugate: Let $f : \mathbf{R}^n \to \mathbf{R}$, then (i)$f(x) \geq f^{**}(x) \forall x \in \mathbf{R}^n$ (ii)If f is closed($\mathbf{epi} f$ is closed) and convex, then $f^{**} = f$

4 Optimization problems

4.1 Standard form

$$\min_x f_0(x)$$

s.t.$f_i(x) \leq 0\ \forall i = 1, \ldots, m$
$h_i(x) = 0\ \forall i = 1, \ldots, p$ f_0 is the *objective* function. f_i is the i^{th} inequality constraint, and h_i is the i^{th} equality constraint. **Domain** of the problem is defined as $D = (\bigcap \mathbf{dom} f_i) \cap (\bigcap \mathbf{dom} h_i)$

4.2 Local optimality

x is *locally optimal* if $\exists R > 0$ such that $f_0(x) = \inf\{f_0(z) : z \in C, |z - x|_2 \leq R\}$, geometrically it means that x minimizes f_0 over its R-neighborhood. If x is feasible and $f_i(x) = 0$, we say the i^{th} inequality constraint is active, otherwise inactive.

4.3 Epigraph form

$$\min_{x,t} t$$

s.t. $t \geq f_0(x)$
$f_i(x) \leq 0\ \forall i = 1, \ldots, m$
$h_i(x) = 0\ \forall i = 1, \ldots, p$
It is equivalent to the original problem.
(x, t) is optimal for the transformed problem $\iff$ x is optimal for the original problem and $t = f_0(x)$

5 Convex optimization

5.1 Standard form

The form follows from the general optimization problem, with additional requirements:

1. f_0 must be convex(if quasiconvex, then the problem is quasiconvex optimization)
2. f_i must be convex.
3. h_i must be affine, that is, $h_i(x) = a_i^T x - b_i$

5.2 Optimality

For convex optimization problems, we have that any local minimum is a global minimum. The optimal set is convex. If the objective is strictly convex, then the optimal set contains at most one point.

5.3 Optimality condition for differentiable f_0

Theorem: Let C be the feasible set for a convex problem. x is optimal $\iff$ $x \in C$ and $\nabla f_0(x)^T (y - x) \geq 0\ \forall y \in C$.
Geometrically, $-\nabla f_0(x)$ defines a supporting hyperplane to the feasible set C, $-\nabla f_0(x)$ makes an obtuse angle with $y - x$. If problem is unconstrained, then the optimal condition reduces to $\nabla f_0(x) = 0$.

6 Duality

6.1 Lagrange multipliers

We consider an idea of relaxing the problem: change the original constrained problem to an unconstrained problem, with additional terms penalizing violation of constraints.
The **Lagrangian** of the primal is defined as

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

where $\lambda \in \mathbf{R}_+^m$, $\mu \in \mathbf{R}^p$.
We call λ, μ the *Lagrangian multipliers* associated with inequality/equality constraints respectively.
The **Lagrange dual function** of the primal is defined as

$$g(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu)$$

g is concave since it is the pointwise infimum of a set of affine functions.

6.2 Weak duality

It can be established that $\forall \lambda, \mu$, $g(\lambda, \mu) \leq p^*$
We define the *dual optimal* as

$$d^* = \max_{\lambda, \mu} g(\lambda, \mu)$$

Weak duality states that $d^* \leq p^*$, as a natural result from the claim above.

6.3 Lagrange dual problem

Consider

$$\max_{\lambda, \mu} g(\lambda, \mu)$$

It is called the **Lagrange dual problem** of the primal problem.
(λ, μ) is *dual feasible* if $\lambda \geq 0$ and $\mu \in \mathbf{R}^p$
(λ^*, μ^*) is *dual optimal* if $g(\lambda^*, \mu^*) = d^*$.
Note: Dual problem is always convex, even if the primal is not. Weak duality always holds, even if primal is not convex.

6.4 Slater's condition

Slater's condition is satisfied when $\exists x' \in \mathbf{relint}(D)$, such that $f_i(x') < 0\ \forall i = 1, \ldots, m, Ax' = b$.
A corollary states that affine inequality constraints need not be strictly satisfied.

6.5 Strong duality

Strong duality states that when we have Slater's condition and the problem is convex, $d^* = p^*$.
Note that Slater's condition is not necessary to establish strong duality.

6.6 Optimality conditions

By weak duality, $g(\lambda,\mu) \le p^*$ for any feasible λ,μ, therefore, given a primal decision variable x, we also know how sub-optimal it is without knowing p^*, by the inequality $f_0(x) - p^* \le f_0(x) - g(\lambda,\mu)$. If we have a sequence of $x^{(k)}$ and $(\lambda^{(k)},\mu^{(k)})$, then we could develop a 'primal-dual' algorithm to obtain x that is arbitrary ϵ-optimal to p^*.

6.7 KKT conditions

Assume an optimization problem with f_i, h_i all differentiable, not necessarily convex. We state the following four conditions:
Stationarity: $\nabla_x L(x,\lambda^*,\mu^*) = 0$, which expands to $\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(x^*) = 0$
Primal feasibility: $f_i(x^*) \le 0\ \forall i = 1,\dots,m$
$h_j(x^*) = 0\ \forall j = 1,\dots,p$
Dual feasibility: $\lambda_i \ge 0\ \forall i = 1,\dots,m$
Complementary slackness: $\lambda_i^* f_i(x^*) = 0\ \forall i = 1,\dots,m$
Any optimization problem described above, with strong duality held, for any primal-dual optimal points (x^*,λ^*,μ^*), must satisfy the KKT conditions. KKT conditions are necessary for asserting primal-dual optimality. If the optimization problem is convex, then KKT conditions are also sufficient.
Therefore, for a differentiable convex optimization problem, the following statements are equivalent:
(i)(x,λ,μ) satisfies KKT conditions.
(ii)Strong duality holds. (x,λ,μ) are primal-dual optimal.
If the problem satisfies Slater's condition, the 'Strong duality holds' part of (ii) could be ignored.

7 Statistical application

7.1 Likelihood interpretation

We consider linear model with independent and identical noise, which is interpreted by the equation

$$y_i = a_i^T x + v_i, i = 1,\dots,m$$

where y_i denote observations, a_i denote feature vectors, v_i denote measurement errors, which is interpreted as noise from i.i.d. distributions $p_v(.)$, and x denote the parameter vector, which is the one we are going to optimize for machine learning process.
The intuition of likelihood is to maximise the probability of, given data y, having such noises v_i to occur across all possible parameter vectors.
The *likelihood* of data y is defined as

$$p_x(y) = \prod_{i=1}^m p_v(y_i - a_i^T x)$$

Equivalently, we can maximize *log-likelihood*, which is defined as

$$l_x(y) = \log p_x(y) = \sum_{i=1}^m \log p_v(y_i - a_i^T x)$$

The best parameter is the optimal solution for

$$\max_{x\in\mathbf{R}^n} \sum_{i=1}^m \log p_v(y_i - a_i^T x)$$

8 Unconstrained minimization

8.1 Problem and insights

$$\min_{x\in\mathbf{R}^n} f(x)$$

such that $f(x) : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable.
Note that in this case, $\nabla f(x^*) = 0$ is necessary and sufficient to establish optimality.
We propose to design an iterative algorithm $x^k \in \mathbf{dom} f$ such that the sequence converges to primal optimal. In practice, we could terminate the algorithm when either $f(x^{(k)}) - p^* \le \epsilon$ or $||\nabla f(x^k)|| \le \epsilon$ for arbitrary ϵ.

8.2 Examples

Quadratic programming: $\frac{1}{2}x^T P x + q^T x + r$, $P \in S_+^n$
We have analytical solution $x^* = -P^{-1}q$
Geometric programming:
$f(x) = \log \sum_{i=1}^m exp(a_i^T x + b_i)$
Too bad, no analytical solution!

8.3 Properties to assume

Strong convexity: We call f is *m-strongly convex* on $S = \{x : f(x) \le f(x^{(0)})\}$, if $\exists m > 0, \nabla^2 f(x) \ge mI\ \forall x \in S$. If f is *m-strongly convex*, then function $g(x) = f(x) - \frac{m}{2}||x||^2$ is convex.
PL-inequality: If f is *m-strongly convex*, $p^* \ge f(x) - \frac{1}{2m}||\nabla f(x)||^2$, which is a derivative from Taylor's theorem: $\forall x, y \in S, \exists z \in S$ on the line segment connecting x and y, such that $f(y) = f(x) + (y-x)^T\nabla f(x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$. Note that PL-inequality implies that when $\nabla f(x)$ is small, then $f(x)$ is close to optimal.
Smoothness: We call f is *M-smooth* if $\exists M > 0$ such that $\nabla^2 f(x) \le MI\ \forall x \in S$
Descent lemma: If f is *M-smooth*, then $f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}||y-x||^2$, which is again a derivative from Taylor's theorem. A consequence of Descent lemma is that $p^x \le f(x) - \frac{1}{2M}||\nabla f(x)||^2$, which is a counterpart to PL-inequality.
Condition number: The ratio $\kappa = \frac{M}{m}$

8.4 Descent method preliminary

We wish to produce a sequence iteratively converging to optimal such that

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

We call t *step size*, and Δx *search direction*.
The search direction must satisfy

$$\nabla f(x^{(k)})^T(\Delta x^{(k)}) < 0$$

such a direction is known as a *descent direction*.

8.4.1 General descent method iteration

Given a starting point $x^{(0)} \in \mathbf{dom} f$

1. Determine a descent direction Δx. For example, the fastest decrease direction is $-\nabla f(x^{(k)})$.
2. Line search: Choose a step size t
3. Update $x^{(k+1)}$.

until termination criterion is met.

8.4.2 Line search

Exact line search: $t = \text{argmin}_{s\ge 0} f(x + s\Delta x)$
Backtracking line search:(Armijo rule)
Choose $\alpha \in (0, 0.5)$, $\beta \in (0,1)$,
while $f(x + t\Delta x) > f(x) + \alpha t\nabla f(x)^T\Delta x$,
update $t = \beta t$

8.5 Gradient Descent

We set search direction $\Delta x = -\nabla f(x)$, and the stopping criterion $||\nabla f(x)|| \le \eta$
Theorem: If gradient descent is implemented on a function f that is m-strongly convex and M-smooth, with exact line search, then the search sequence converges to optimal in linear rate:

$$f(x^{(k)}) - p^* \le c^k(f(x^{(0)}) - p^*)$$

where $c = 1 - \kappa$.
Theorem: If f is m-strongly convex and M-smooth, by backtracking line search with parameter (α,β), then

$$f(x^{(k)}) - p^* \le d^k(f(x^{(0)}) - p^*)$$

where $d = 1 - \min(2m\alpha, 2\beta\alpha\frac{m}{M})$

8.6 Steepest descent

We could approximate $f(x+v) \approx f(x) + \nabla f(x)^T v$, and an intuition is to choose v to make $\nabla f(x)^T v$ as negative as possible. We define normalized steepest descent $\Delta x_{nsd} = \text{argmin}(\nabla f(x)^T v : ||v|| = 1)$, with a chosen norm. When Euclidean norm is considered, we recover gradient descent.

8.7 Stochastic gradient descent

8.7.1 Preliminary

Based on a set of data and observations $(x_i, y_i), i \in [m]$, and a proposed model with parameter θ, we minimize a chose loss function, defined by $f(x) = \text{loss}(\theta, D) = \frac{1}{m}\sum_{i=1^m} \text{loss}(\theta, x_i, y_i)$
The only difference with typical gradient descent is that now we choose a subset of data points, instead of using all m points to compute gradient to optimize model parameters.
The problem is formulated as

$$\min_{x\in\mathbf{R}^n} f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)$$

8.7.2 Assumptions

1. $p^* > -\infty$, the loss is bounded below.
2. loss function f is M-smooth
3. $\sup_{x\in\mathbf{R}^n} E[||\nabla f_i(x)||^2] \le \sigma^2$, variance bound.

8.7.3 Efficiency

Theorem: If we run stochastic gradient descent for T iterations, we have

$$\min_{k\in[T]} E[||\nabla f_i(x)||^2] \le \frac{f(x^{(0)}) - p^*}{\sum_{k=0}^{T-1} t_k} + \frac{M\sigma^2}{2}\frac{\sum_{k=0}^{T-1} t_k^2}{\sum_{k=0}^{T-1} t_k}$$

If we choose constant step size $t_k = t$, and assume $\sigma^2 = 0$, then RHS reduces to $\frac{f(x^{(0)}) - p^*}{tT}$, thus we have a convergence rate of $O(\frac{1}{T})$. We could see that without assuming m-strongly convex, the convergence is much slower.

8.8 Newton's method

8.8.1 Preliminary

Hessian norm of matrix v associated with matrix P : $||v||_P = \sqrt{v^T P v}$
Newton's direction: $\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$
We could show that Δx_{nt} is a descent direction, since $\Delta x_{nt}^T \nabla f(x) = (-\nabla^2 f(x)^{-1}\nabla f(x))^T \nabla f(x) = -\nabla f(x)^T \nabla^2 f(x)^{-1}\nabla f(x) < 0$
From now on, we further assume that $\nabla^2 f(x)$ and $\nabla^2 f(x)^{-1}$ are positive definite.
We could observe that when Hessian norm associated with $\nabla^2 f(x)$ is chosen as the norm for steepest descent, we obtain Newton's direction. Since the direction is obtained by

$$\Delta x_{nt} = \text{argmin}_v\{\nabla f(x)^T v : v^T \nabla^2 f(x) v = 1\}$$

Solve the problem by KKT conditions will yield the desired result.
Alternatively, if we consider the quadratic Taylor approximation of f at x, then

$$f(x+v) \approx f(x) + \nabla f(x)^T v + \frac{1}{2}v^T\nabla^2 f(x) v$$

Differentiate with respect to v will yield the desired result.

8.8.2 Algorithm

Given a starting point $x^{(0)} \in \mathbf{dom} f$, and $\epsilon > 0$.

1. Compute Δx_{nt}
2. Compute Newton decrement $\lambda(x) = ||\nabla f(x)||_{\nabla^2 f(x)^{-1}}$
3. Terminate iteration if $\frac{\lambda(x)^2}{2} < \epsilon$
4. Line search and update next step. Note that when x is close to optimum, $t = 1$

8.8.3 Efficiency

Theorem: We consider $g : \mathbf{R}^n \to \mathbf{R}^n$, and ∇g is symmetric Jacobian matrix.
For $\delta > 0$, define $S_\delta = \{x \in \mathbf{R}^n : ||x - x^*|| \le \delta\}$. Assume g is continuously differentiable in S_δ and ∇g is invertible.
If $\exists \delta > 0$ such that $x^{(0)} \in S_\delta$, the sequence $\{x^{(k)}\}$ converges to x^* and the convergence is superlinear, which means that $\lim_{k\to\infty} \frac{||x^{(k+1)} - x^*||}{||x^{(k)} - x^*||} = 0$
If $\exists L, M > 0$ such that $\forall x, y \in \mathbf{dom} g$, we have $||\nabla g(x) - \nabla g(y)|| \le L||x-y||$, and $||\nabla g(x)^{-1}|| \le M$, then if $x^{(0)} \in S_\delta$, we have $||x^{(k+1)} - x^*|| \le \frac{LM}{2}||x^{(k)} - x^*||^2$. It implies that the convergence rate of Newton's method is quadratic.