DSA3102: Lecture 12 (Readings Sec 9.4, 9.5 & my notes on Newton's method)

↳ very fast under the right conditions

Last time: An optimization algorithm produces a sequence $\{x^{(k)}\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$ with the hope that $x^{(k)} \to x^*$

(infinite collection of vectors)

$x^* = \operatorname{argmin}_x f(x)$. If $f(x)$ strictly convex, x^* is unique

start at x^0

Gradient descent: $x^{(k+1)} = x^{(k)} - t \nabla f(x^{(k)})$ move in direction of negative gradient

Stepsize choice: $t = \operatorname{argmin}_{s \geq 0} f(x^{(k)} - s \nabla f(x^{(k)}))$

Backtracking line search.

↳ only make use of first derivative

Conditions on f: (Hessian)

$m < M$

① m-strongly convex: $\nabla^2 f(x) \geq mI \quad \forall x \in S$.

② M-smoothness: $\nabla^2 f(x) \leq MI \quad \forall x \in S$.

Thm: GD with stepsize chosen according to exact line search on f that satisfies m-s.c. & M-smoothness yields.

linear convergence

k^{th} iteration

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{m}{M}\right)^k \left(f(x^{(0)}) - p^*\right)$$

sub-optimality at the start

geometric series ↳ go down exponentially fast

$$\|x^{(k)} - x^*\| \leq c \|x^{(k-1)} - x^*\|$$

$0.8 \times 10^{-6} \to 0.64 \times 10^{-6}$

$c = 0.8$, $\|x^{(k-1)} - x^*\| = 10^{-6}$

Thm: GD with stepsize chosen according to exact line search on f that satisfies M-smoothness yields

(no m strong convex)

cannot guarantee you converge so quickly

$f(x^{(k)}) - p^* = O(1/k)$ goes down Polynomially fast

Snapshot of today main theoretical result:

$\|x^{(k+1)} - x^*\| \leq c \|x^{(k)} - x^*\|^2$ quadratic convergence

---

Recall Steepest Descent: motivation for newton's method

perturbation

$f(x+v) \cong f(x) + \nabla f(x)^T v$

deviation

Minimize $\nabla f(x)^T v$ over all directions v. (want to push F to be as small as possible)

Def: Normalized steepest descent direction.

$\Delta x_{nsd} = \text{argmin}\{\nabla f(x)^T v : \|v\| = 1\}$

norm 1

Def: Unnormalized steepest descent direction.

$\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$.

dual norm

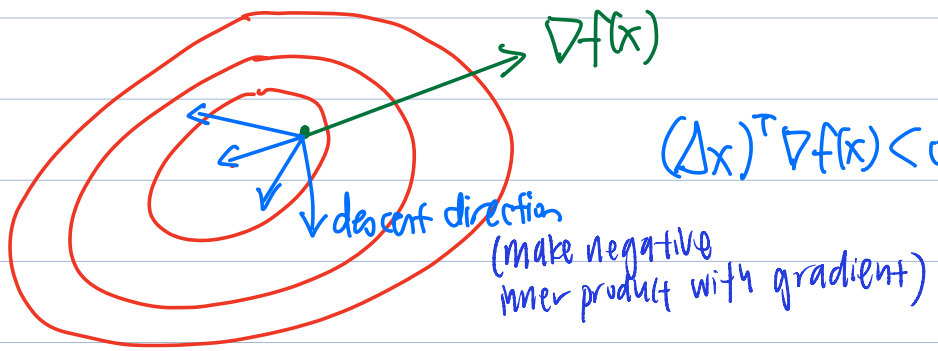
Dual norm: $\|x\|_* := \sup\{y^T x : \|y\| \leq 1\}$

$-\|x\|_* := \inf\{y^T x : \|y\| \leq 1\}$

Rmk: For the steepest descent step,

consider inner product between gradient and steepest descent direction

$$\nabla f(x)^T \Delta x_{sd} = \nabla f(x)^T \left[\|\nabla f(x)\|_* \Delta x_{nsd} \right]$$
$$= \|\nabla f(x)\|_* \nabla f(x)^T \Delta x_{nsd}$$
$$= \|\nabla f(x)\|_* \left(-\|\nabla f(x)\|_* \right)$$
$$= -\|\nabla f(x)\|_*^2 < 0$$



The steepest descent direction is a descent direction.

$$-\|a\|_2 \|b\|_2 \le a^T b \le \|a\|_2 \|b\|_2$$ Cauchy

Ex: Euclidean norm.

$$\Delta x_{nsd} = \text{argmin} \{ \nabla f(x)^T v : \|v\|_2 = 1 \}.$$
$$= -\nabla f(x)$$ use lower bound of cauchy inequality

different norm get different algorithm

Rmk: If we instantiate steepest descent to the Euclidean norm, we get gradient descent.

$\frac{1}{p} + \frac{1}{q} = 1$

Ex: l_1-norm $\quad -\|a\|_1 \|b\|_\infty \le a^T b \le \|a\|_1 \|b\|_\infty$ (p, q | p, q)

$$\Delta x_{nsd} = \text{argmin} \{ \nabla f(x)^T v : \|v\|_1 = 1 \}.$$

$\nabla f(x)^T v = -|\nabla f(x)|_\infty$

Differentiate wrt one coordinate

By holder inequality, one vector constrained to l_1 norm, the other should be constrained by ∞ norm

$\|y\|_\infty = \max_{1 \le i \le n} |y_i|$ $\quad\quad$ $\|y\|_1 = \sum_{i=1}^{n} |y_i|$

Let $i \in [n]$ be s.t. $|(\nabla f(x))_i| = \|\nabla f(x)\|_\infty$ — partial derivative in ith coordinate

$\Delta x_{nsd} = -\text{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i$.

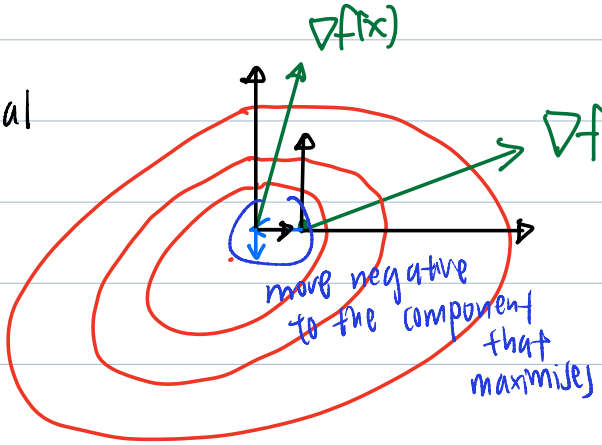
either up/down

look for coordinate which has maximum abs value

$e_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i$th position (move in only 1 particular direction)

Rmk: If we instantiate steepest descent to the l_1 norm, we get coordinate descent, i.e., move in the direction of maximal decrease.

find direction which yields maximal decrease



can only move coordinate wise

only move in axis-aligned direction

Newton's Method

Newton direction: $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$. (hessian)

Descent direction: $(\Delta x_{nt})^T \nabla f(x) = (-\nabla^2 f(x)^{-1} \nabla f(x))^T \nabla f(x)$

take inner product of presumed desc direction with gradient and check that it is -ve

must be positive definite

$$= -\nabla f(x)^T \boxed{\nabla^2 f(x)^{-1}} \nabla f(x) < 0.$$

The Newton direction is a descent direction.

Rmk: We assume $\nabla^2 f(x)$ & $\nabla^2 f(x)^{-1}$ are positive definite. (so that it is invertible)

Ex: $f(x) = \frac{1}{2} x^T Q x \qquad Q = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}.$

$= \frac{1}{2}[0.01 x_1^2 + x_2^2].$

$x^* = \underline{0} \qquad -\nabla f(x) = -Qx = \begin{bmatrix} 0.01 x_1 \\ x_2 \end{bmatrix}.$

If we start GD at $x^{(0)}$, the update is

$$x^{(1)} = x^{(0)} + t \begin{bmatrix} -0.01 x_1^{(0)} \\ -x_2^{(0)} \end{bmatrix}.$$

Is this a good direction?

What is the best direction if we start at $x^{(0)} \neq \underline{0}$?

It is $-x^{(0)} = \begin{bmatrix} -x_1^{(0)} \\ -x_2^{(0)} \end{bmatrix}$ ← a direction

Compare the GD direction $\begin{bmatrix} -0.01 x_1^{(0)} \\ -x_2^{(0)} \end{bmatrix}$ with the best direction $\begin{bmatrix} -x_1^{(0)} \\ -x_2^{(0)} \end{bmatrix}$.

→ get to the optimum straightaway

They can be very diff. due to coefficients

Newton direction $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$ converge in one single step

$= -Q^{-1} Qx = -x.$

$f(x) = \frac{1}{2} x^T Q x$

$\nabla f(x) = Qx$

$\nabla^2 f(x) = Q$ invertible

↑ gets best direction (the minus of the point)

Newton direction: $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$

Steepest descent: $\Delta x_{nsd} = \text{argmin} \{ \nabla f(x)^T v : \|v\|_{..} = 1 \}$

$\Delta x_{sd} = -\|\nabla f(x)\|_* \Delta x_{nsd}.$

$\Delta x_{nt} = \text{argmin} \{ \nabla f(x)^T v : \|v\|_{\nabla^2 f(x)} = 1 \}$ ← Hessian norm

steepest desc with hessian norm

$\|v\|_P = \sqrt{v^T P v}$

Verify that Δx_{sd} when instantiated to the Hessian norm gives Newton's direction.

$c = \nabla f(x)$ $P = \nabla^2 f(x) > 0.$

$\min_v \{ c^T v : v^T P v = 1 \}.$ $\min_v \{ \nabla f(x) v : v^T \nabla^2 f(x) v = 1 \}$

KKT condition: $L(v, \mu) = c^T v + \mu (v^T P v - 1)$

nodual
feasibility as no inequality constraints

derivative (wrt v

Stationarity: $\nabla_v L(v,\mu) = c + 2\mu P v = 0$

$$v = -\frac{1}{2\mu} P^{-1} c.$$

Primal feasibility: $v^T P v = 1$

$$\left(-\frac{1}{2\mu} P^{-1} c\right)^T P \left(-\frac{1}{2\mu} P^{-1} c\right) = 1$$

$$\mu^2 = \frac{1}{4} c^T P^{-1} c \Rightarrow v = -(c^T P^{-1} c)^{-1/2} P^{-1} c.$$

$$\Delta x_{nsd} = -\left((\nabla f(x))^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{-1/2} \nabla^2 f(x)^{-1} \nabla f(x).$$

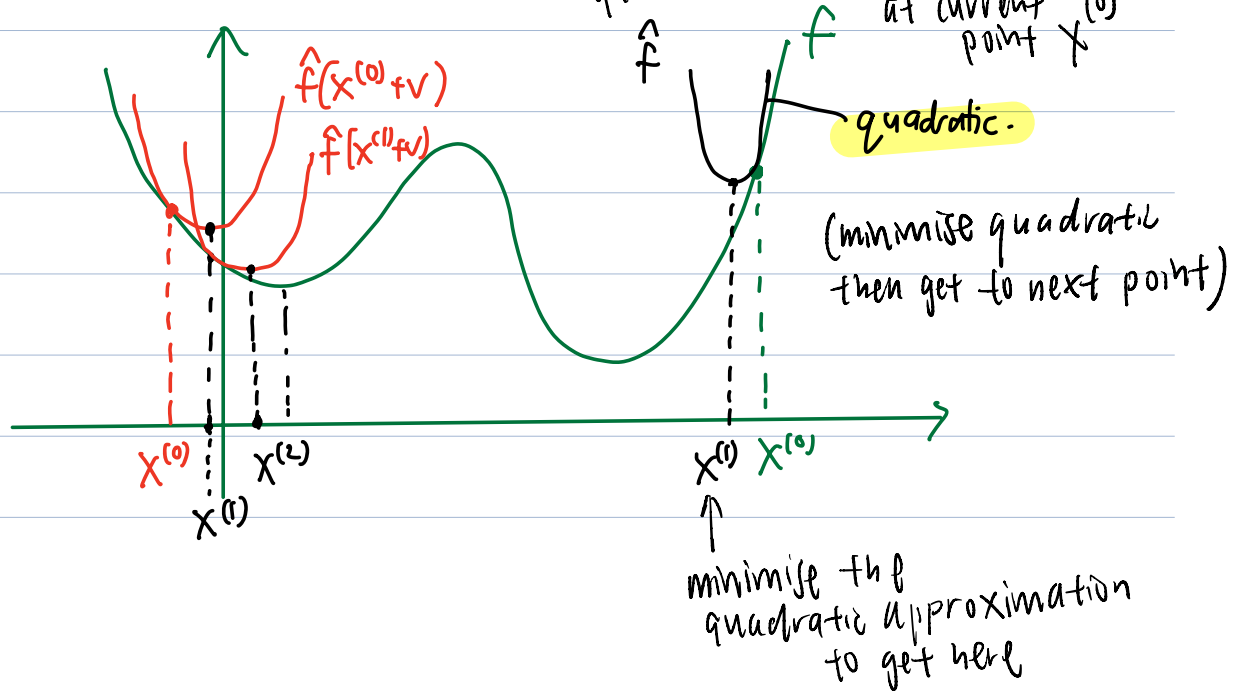
$$\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

Minimizer of the second-order approximation to f

2nd order Taylor series

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

quad-approx



newton direction → best direction according to 2nd order Taylor approx of f at current point

Differentiate $f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$ w.r.t. v.

$$\nabla f(x) + \nabla^2 f(x) v = 0$$

direction to move

$$v = -\nabla^2 f(x)^{-1} \nabla f(x).$$

Solution of linearized optimality condition.

$g = \nabla f$ gradient of f $\quad g(x^*) = 0 \quad \because \nabla f(x^*) = 0.$

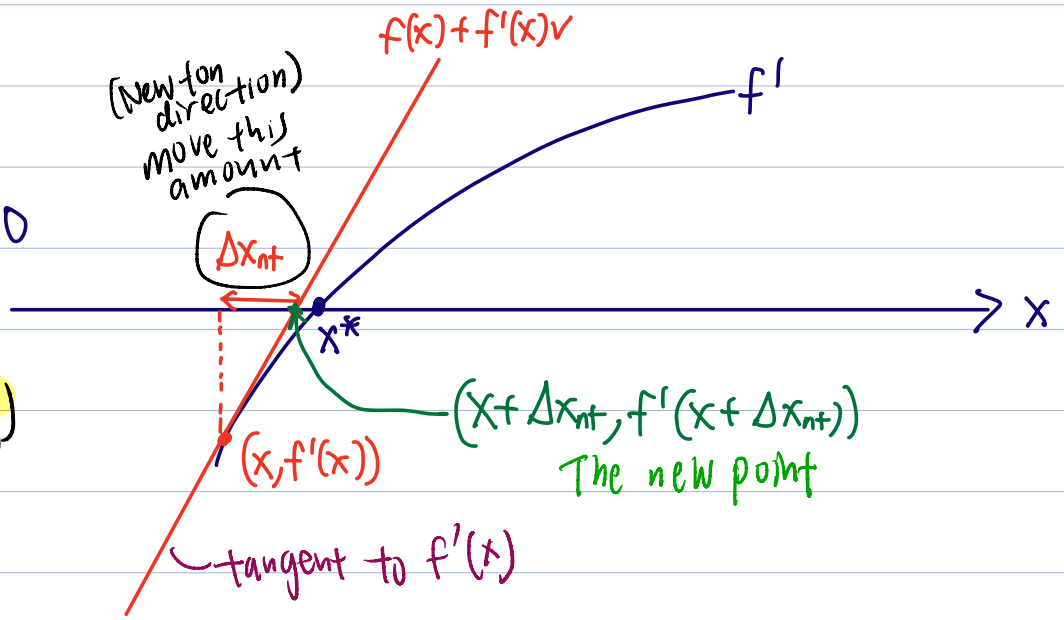
Linearizing this near x: $\nabla f(x) + \nabla^2 f(x) v = 0$

$$v = -\nabla^2 f(x)^{-1} \nabla f(x)$$

$n = 1.$

$f'(x^*) = 0$

want to find x^* (zero crossing)



Affine invariance of Newton step:

Newton step is independent of linear or affine changes of the coordinates.

T: matrix
x, y: vector

$T \in \mathbb{R}^{n \times n}$: non-singular. (invertible)

$x, y \in \mathbb{R}^n$

$g(y) = f(Ty)$ $\qquad x = Ty.$

$$\nabla g(y) = T^T \nabla f(x), \qquad \nabla^2 g(y) = T^T \nabla^2 f(x) T.$$

$$\begin{aligned}
\Delta y_{nt} &= -\nabla^2 g(y)^{-1} \nabla g(y) \\
&= -(T^T \nabla^2 f(x) T)^{-1} T^T \nabla f(x) \\
&= -T^{-1} \nabla^2 f(x)^{-1} \cancel{T^{-T}} \cancel{T^T} \nabla f(x) \\
&= -T^{-1} \underbrace{\nabla^2 f(x)^{-1} \nabla f(x)}_{\text{newton step in the original coordinate system}} \\
&= T^{-1} \Delta x_{nt}
\end{aligned}$$

$\Rightarrow$ Newton steps of f and g are related by the same linear transformation.

$$x + \Delta x_{nt} = T(y + \Delta y_{nt})$$

Newton's Method:

Given a starting point $x^{(0)} \in \text{dom } f$. Tol $\varepsilon > 0$ (Tolerance)

repeat (loop)

1. Compute newton direction $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$ ← computationally costly.

   Newton decrement: $\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{1/2}$

   ↳ smaller this is, the closer to the optimum

Problem of Newton method:
inverting a matrix takes (step 1)
long time → Solution: Approximate the Hessian (diagonal matrix)
(not scalable)

$\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}}$

2. Quit if $\lambda(x)^2/2 < \varepsilon$
3. Line search: Choose t by backtracking line search or other methods.
4. $x^+ = x + t\Delta x_{nt}$.
   step size. newton direction

Analysis of Newton's method near the optimum.

converge very quickly (pure Newton phase)

$$x^{(k+1)} = x^{(k)} - t\nabla^2 f(x^{(k)})^{-1}\nabla f(x)$$

Near the optimum, backtracking line search results in no reduction of step size $\Rightarrow t = 1$.
step size always 1

Can be seen as a solver for the system of n equations.

$$g(x) = 0 \qquad g: \mathbb{R}^n \to \mathbb{R}^n$$

find x such that derivative vanishes

$g = \nabla f$

$\nabla g = \nabla^2 f$.

(find $x^{(k)} \to x^*$)

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

Intuitive argument to show why Newton converges very fast to x^* satisfying $g(x^*) = 0$.

Taylor series approx

$$0 = g(x^*) = g(x^{(k)}) + \nabla g(x^{(k)})^T (x^{(k)} - x^*) + o(\|x^{(k)} - x^*\|)$$

smaller terms

Multiply this with $\nabla g(x^{(k)})^{-1}$

$$x^{(k)} - x^* - \nabla g(x^{(k)})^{-1} g(x^{(k)}) = o(\|x^{(k)} - x^*\|).$$

$$x^{(k+1)} - x^* = o(\|x^{(k)} - x^*\|)$$ ← small thing

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = 0.$$ ← superlinear convergence.

the dist from next iterate to optimum divided by dist from current iterate to optimum is 0

Compare this to GD $\|x^{(k+1)} - x^*\| \le 0.8 \|x^{(k)} - x^*\|$

$$\limsup_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} \le 0.8$$ $1 - \frac{m}{M}$

⇒ linear convergence

Thm: $g: \mathbb{R}^n \to \mathbb{R}^n$ ∇g is symmetric Jacobian matrix.
(not examinable)

For $\delta > 0$, let $S_\delta = \{x \in \mathbb{R}^n : \|x - x^*\| \le \delta\}$.
↳ neighbourhood of optimum (very close to x^*)

Assume g is continuously differentiable in S_δ and that $\nabla g(x^*) \in \mathbb{R}^{n\times n}$ is invertible.

1. $\exists \delta > 0$ s.t. if $x^{(0)} \in S_\delta$ (initialise very close to optimum), the sequence $\{x^{(k)}\}$ converges to x^* and the convergence of

invertible at x^* ⟺ in a neighbourhood also invertible since gradient is continuous

$\{\|x^{(k)} - x^*\|\}$ is super-linear.

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = 0.$$

2. If $\exists L, M > 0$ s.t.

$$\|\nabla g(x) - \nabla g(y)\| \le L\|x - y\| \quad \forall x, y$$
$$\|(\nabla g(x))^{-1}\| \le M$$

then if $x^{(0)} \in S_\delta$, we have

← at $(k+1)$-th step, how far away from optimum

$$\|x^{(k+1)} - x^*\| \le \frac{LM}{2}\|x^{(k)} - x^*\|^2$$

Convergence is quadratic.

Pf: Choose $\delta > 0$ s.t. $(\nabla g(x))^{-1}$ exists for all $x \in S_\delta$.

Claim: $g(x^{(k)}) = \int_0^1 \nabla g(x^* + t(x^{(k)} - x^*))^T (x^{(k)} - x^*)\, dt$.

Pf: Let $h(t) = g(x^* + t(x^{(k)} - x^*))$

$$h(1) - h(0) = \int_0^1 \frac{dh}{dt}\, dt \qquad -(1)$$

$h(1) = g(x^{(k)})$, $h(0) = g(x^*)$

By the chain rule applied to h,

$$\frac{dh}{dt} = \nabla g(x^* + t(x^{(k)} - x^*))^T (x^{(k)} - x^*) \qquad -(2)$$

$$g(x^{(k)}) = \int_0^1 \nabla g(x^* + t(x^{(k)} - x^*))^T (x^{(k)} - x^*)\, dt.$$

as desired.

Now we will estimate $\|x^{(k+1)} - x^*\|$ in terms of $\|x^{(k)} - x^*\|$. (provide upper bound)

$$\|x^{(k+1)} - x^*\| = \left\| x^{(k)} - \nabla g(x^{(k)})^{-1} g(x^{(k)}) - x^* \right\|$$

$$= \left\| x^{(k)} - x^* - \nabla g(x^{(k)})^{-1} g(x^{(k)}) \right\|$$

$$= \left\| \nabla g(x^{(k)})^{-1} \left(\nabla g(x^{(k)})(x^{(k)} - x^*) - g(x^{(k)}) \right) \right\|$$

$$\overset{\text{Claim}}{=} \left\| \nabla g(x^{(k)})^{-1} \left[\nabla g(x^{(k)}) - \int_0^1 \nabla g(x^* - t(x^{(k)} - x^*))^T dt \right] (x^{(k)} - x^*) \right\|$$

$$= \left\| \nabla g(x^{(k)})^{-1} \left[\int_0^1 \nabla g(x^{(k)}) - \nabla g(x^* - t(x^{(k)} - x^*))^T dt \right] (x^{(k)} - x^*) \right\|$$

$$\leq \|\nabla g(x^{(k)})^{-1}\| \left\| \int_0^1 \nabla g(x^{(k)}) - \nabla g(x^* - t(x^{(k)} - x^*))^T dt \right\| \cdot \|x^{(k)} - x^*\|$$

$$\leq M \underbrace{\left\| \int_0^1 \nabla g(x^{(k)}) - \nabla g(x^* - t(x^{(k)} - x^*))\, dt \right\|}_{A < \varepsilon} \|x^{(k)} - x^*\|$$

By the continuity of ∇g, we may take δ suff. small to make A arbitrarily small.

$\Rightarrow$ $x^{(k)}$ converges x^*

& $\{\|x^{(k)} - x^*\|\}$ converges superlinearly fast

$\because$ A can be made arbitrarily small.

$\|\nabla g(x) - \nabla g(y)\| \le L\|x-y\|$

Part 2: $\|x^{(k+1)} - x^*\| \le M A \|x^{(k)} - x^*\|$

$$A = \left\| \int_0^1 \nabla g(x^{(k)}) - \nabla g(x^* + t(x^{(k)} - x^*))\, dt \right\|$$

$$\le \int_0^1 \| \nabla g(\underbrace{x^{(k)}}_{x}) - \nabla g(\underbrace{x^* + t(x^{(k)} - x^*)}_{y}) \|\, dt.$$

$$\le L \int_0^1 \| x^{(k)} - (x^* + t(x^{(k)} - x^*)) \|\, dt.$$

$(1-t)(x^{(k)} - x^*)$

$$= L \int_0^1 (1-t) \|x^{(k)} - x^*\|\, dt.$$

$$= L \|x^{(k)} - x^*\| \underbrace{\int_0^1 1 - t\, dt}_{1/2}.$$

$$A \le \frac{L}{2} \|x^{(k)} - x^*\|$$

$$\|x^{(k+1)} - x^*\| \le M A \|x^{(k)} - x^*\|$$

$$\le \frac{LM}{2} \|x^{(k)} - x^*\|^2$$ (qed) ////.