CHAPTER #9

PROBLEM OF THIS CHAPTER: FOR A CONVEX $f(x)$,

minimize: $f(x)$ i.e. NO CONSTRAINTS. ALSO,

$f(x)$ IS TWICE CONTINUOUSLY DIFFERENTIABLE ($\Rightarrow$ dom f IS OPEN)

NECESSARY AND SUFFICIENT CONDITION FOR A POINT TO BE OPTIMAL: $\nabla f(x^*) = 0$, WHICH IS SYSTEM OF n EQUATIONS WITH n UNKNOWNS. HOWEVER, WE CAN RARELY SOLVE IT IN PRACTICE. SO WE FIND SEQUENCE

$x^{(k)}$ WITH $f(x^{(k)}) \to p^*$ AS $k \to \infty$

MINIMIZING SEQUENCE AND STOP WHEN $f(x^{(k)}) - p^* \leq \epsilon$

REQUIREMENT: WE NEED A STARTING POINT $x^{(0)} \in \text{dom} f$ SUCH THAT THE SUBLEVEL SET

$$S = \{ x \in \text{dom} f \mid f(x) \leq f(x^{(0)}) \}$$ IS CLOSED

(OTHERWISE, THE ABOVE SEQUENCE MAY CONVERGE TO A POINT PT IN S). THIS IS GUARANTEED IF f IS CLOSED

DEFINITION: f IS CLOSED IF ALL ITS SUBLEVEL SETS ARE CLOSED

PROPERTY: 1) CONTINUOUS FUNCTIONS WITH CLOSED DOMAINS (SUCH AS $\mathbb{R}^n$) ARE CLOSED

2) CONTINUOUS FUNCTIONS WITH OPEN DOMAINS FOR WHICH $f(x)$ TENDS TO INFINITY AS $x \to$ bd dom f ARE CLOSED.

EXAMPLE #1 minimize $\frac{1}{2} x^T P x + q^T x + r$,

$P \in S_+^n$, $q \in \mathbb{R}^n$, $r \in \mathbb{R}$

$\nabla \left(\frac{1}{2} x^T P x + q^T x + r \right) = 0 \Leftrightarrow \boxed{Px + q = 0}$ WHICH ARE LINEAR

CASES: 1) ONE SOLUTION (ALWAYS THE CASE WHEN $P > 0$)
⇒ UNIQUE MINIMUM

2) INFINITE SOLUTIONS (INFINITE MINIMA)

3) NO SOLUTIONS ⇒ UNBOUNDED BELOW.

NOTE THAT SUBLEVEL SET CONDITION IS SATISFIED

EXAMPLE #2 minimize $f(x) = \log\left(\sum_{i=1}^{m} \exp\left[a_i^T x + b_i\right]\right)$

OPTIMALITY CONDITION

$$\frac{\partial f(x)}{\partial x_j} = 0 \Leftrightarrow \frac{1}{\log\left(\sum_{j=1}^{m} \exp\left[a_j x + b_j\right]\right)} \sum_{i=1}^{m} \exp\left[a_i^T x + b\right] a_{ij} = 0$$

THEREFORE, $\nabla f(x) = 0 \Leftrightarrow$

$$\frac{1}{\sum_{j=1}^{m} \exp\left[a_j^T x + b_j\right]} \cdot \sum_{i=1}^{m} \exp\left[a_i^T x + b_i\right] a_i = 0$$

WHICH CANNOT BE SOLVED ANALYTICALLY, SO HERE WE NEED ITERATIVE METHOD. AGAIN, SUBLEVEL SET CONDITION IS SATISFIED

EXAMPLE #3 minimize $f(x) = -\sum_{i=1}^{m} \log\left(b_i - a_i^T x\right)$,

I.E. THE LOGARITHMIC BARRIER FOR THE SET OF THE INEQUALITIES. AGAIN, THE CONDITION IS SATISFIED. REGARDING THE OPTIMALITY CONDITION,

$$\frac{\partial f}{\partial x_j} = 0 \Leftrightarrow -\sum_{i=1}^{m} \frac{1}{b_i - a_i^T x} \cdot \left(-a_{ij}\right) = \sum_{i=1}^{m} \frac{a_{ij}}{b_i - a_i^T x} = 0,$$

WHICH AGAIN CANNOT BE SOLVED ANALYTICALLY

IN VECTOR FORM,

$$\nabla f(x) = 0 \Leftrightarrow \sum_{i=1}^{m} \frac{a_i}{b_i - a_i^T x} = 0$$

AGAIN, SUBLEVEL CONDITION IS SATISFIED

9.1.2 STRONG CONVEXITY

LET f BE TWICE CONTINUOUSLY DIFFERENTIABLE.

DEFINITION: f IS STRONGLY CONVEX ON S IF $\exists m > 0$:

$$(9.7)\quad \nabla^2 f(x) \geq mI \iff \nabla^2 f(x) - mI \geq 0 \quad \forall x \in S.$$

SO LET'S ASSUME THAT f IS STRONGLY CONVEX.

PROPERTY 1: $\forall x, y \in S$,

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|_2^2 \qquad (9.8)$$

($m=0 \Rightarrow$ KNOWN CASE)

Proof:

KNOWN PROPERTY: FOR $x, y \in S$,

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

WHERE z ON THE LINE SEGMENT $[x,y]$. BY (1),

LAST TERM IS AT LEAST $\frac{m}{2}\|x-y\|_2^2$, SO RESULT FOLLOWS

PROPERTY 2: LET ANY $x \in S$. THEN

$$f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2,$$ SO WE BOUND SUBOPTIMALITY

Proof:

WE KNOW

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|_2^2$$

IT'S MINIMIZE THE RIGHT HAND SIDE W.R.T y:

$$\nabla\left[f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}(y-x)^T I (y-x)\right] = 0 \iff$$

$$\nabla f(x) + \frac{m}{2}\cdot 2(y-x) = 0 \iff$$

$\tilde{y} = x - \frac{1}{m}\nabla f(x)$, THEREFORE

$$f(y) \geq f(x) + \nabla f(x)^T [\tilde{y} - x] + \frac{m}{2}\|\tilde{y}-x\|_2^2$$

$$= f(x) + \nabla f(x)^T \left(-\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left(\frac{1}{m^2}\right)\|\nabla f(x)\|_2^2$$

$$= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$$

THIS BOUND HOLD FOR ALL y, SO IT WILL ALSO HOLD FOR $\inf_y f(y)$, SO $p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$

AND THE RESULT FOLLOWS

PROPERTY 3: $\|\nabla f(x)\|_2 \leq (2m\epsilon)^{\frac{1}{2}} \Rightarrow f(x) - p^* \leq \epsilon$

(PROOF COMES IMMEDIATELY FROM PREVIOUS.)

PROPERTY 4: $\|x - x^*\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2$ → (THEREFORE x^* IS UNIQUE)

PROOF:

WE APPLY (9.8) OF PREVIOUS PAGE FOR $y = x^*$

$$\Rightarrow p^* = f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{m}{2}\|x^* - x\|_2^2$$

$$\geq f(x) - \|\nabla f(x)^T\| \|x^* - x\| + \frac{m}{2}\|x^* - x\|_2^2$$

(BY CAUCHY-SCHWARZ), AND BECAUSE $p^* - f(x) \leq 0$, IT FOLLOWS THAT

$$-\|\nabla f(x)^T\| \|x^* - x\| + \frac{m}{2}\|x^* - x\|^2 \leq 0,$$

AND THE RESULT FOLLOWS.

PROPERTY 5: THE HESSIAN IS ALSO UPPER BOUNDED!

IN PARTICULAR ANY SUBLEVEL SET IS BOUNDED, AND BECAUSE THE MAXIMUM EIGENVALUE IS A CONTINUOUS FUNCTION OF x, IT IS ALSO UPPER BOUNDED $\Rightarrow$

$$\exists M: \nabla^2 f(x) \leq MI$$

PROPERTY 6:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}\|y - x\|_2^2 \quad (9.13)$$

Proof: R.H.S is $f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}(y-x)^T I (y-x)$

SET $z = y - x$ AND MINIMIZE W.R.T z:

$$\nabla \left[f(x) + \nabla f(x)^T z + \frac{M}{2} z^T I z \right] = 0 \Leftrightarrow$$

$$\nabla f(x) + Mz = 0 \Leftrightarrow \quad z = -\frac{1}{M} \nabla f(x) \Leftrightarrow$$

$y = x - \frac{1}{M} \nabla f(x)$, SO MINIMUM R.H.S. IS

$$f(x) + \nabla f(x)^T \left(-\frac{1}{M} \nabla f(x) \right) + \frac{M}{2} \left(\frac{1}{M^2} \right) \| \nabla f(x) \|_2^2$$

$$= f(x) - \frac{1}{2M} \| \nabla f(x) \|_2^2$$ IT FOLLOWS THAT $\forall y$,

$$f(y) \leq f(x) - \frac{1}{2M} \| f(x) \|_2^2$$ AND THE RESULT FOLLOWS

CONDITION NUMBER OF SUBLEVEL SETS:

(9.7), (9.12) $\Rightarrow$ $mI \leq \nabla^2 f(x) \leq MI \quad \forall x \in S$

$\Rightarrow$ $K = \frac{M}{m}$ IS UPPER BOUND ON $\text{COND}(\nabla^2 f(x)) \triangleq \frac{\lambda_{max}(\nabla^2 f(x))}{\lambda_{min}(\nabla^2 f(x))}$

K IS VERY IMPORTANT, BECAUSE IT AFFECTS THE SHAPE OF SUBLEVEL SETS, AND, THROUGH THEM, CONVERGENCE RATES OF ALGORITHMS. WE EXPLORE THE SHAPE MORE:

LET $mI \leq f(x) \leq MI \quad \forall x \in S$

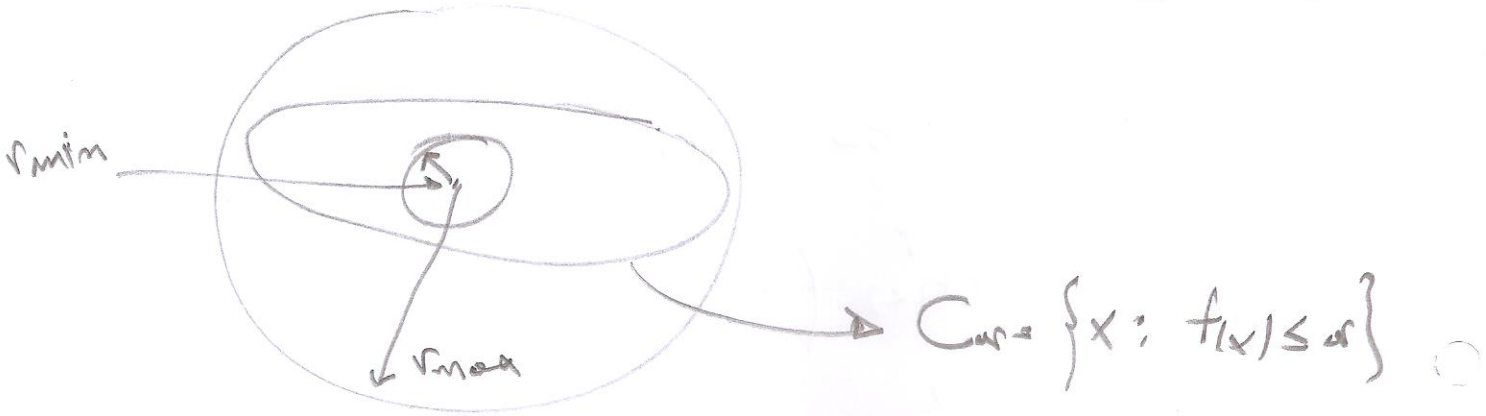
(9.13), (9.8), $x = x^*$ $\Rightarrow$ $p^* + \frac{M}{2} \|y-x\|_2^2 \geq f(y) \geq p^* + \left(\frac{m}{2}\right) \|y-x\|_2^2$

SET $f(y) = \alpha$

$$p^* + \frac{M}{2}\|y-x\|_2^2 \geq \alpha \Rightarrow \|y-x^*\|_2 \geq r_{min} = \left[\frac{2(\alpha-p)}{M}\right]^{\frac{1}{2}}$$

LIKEWISE:

$$p^* + \left(\frac{m}{2}\right)\|y-x^*\|_2^2 \leq \alpha \Leftrightarrow m\|y-x^*\|_2 \leq r_{MAX} = \left[\frac{2(\alpha-p)}{m}\right]^{\frac{1}{2}}$$



SO WHEN $k \simeq 1$ THE SUBLEVEL SET ARE GUARANTEED TO BE APPROXIMATELY CIRCULAR.

TAYLOR SERIES JUSTIFICATION

$$f(y) \simeq p^* + \nabla f(x^*)(y-x^*) + \frac{1}{2}(y-x^*)\nabla^2 f(x^*)(y-x^*)$$

($\nabla f(x^*) \to 0$ AT OPTIMUM x^*)

WHEN $y \simeq p^*$. THEREFORE, CLOSE TO OPTIMUM, EQUIPOTENTIAL LINES LOOK LIKE ELLIPSES WITH ECCENTRICITY DEPENDING OF $\frac{M}{m}$

A LAST CAVEAT: M, m ARE UNKNOWN, MOST OF THE TIME. SO WHAT WE HAVE SEEN IS USEFUL FOR UNDERSTANDING PROBLEMS, AND NOT SO MUCH FOR SOLVING THEM.

9.2 DESCENT METHODS

ALL ALGORITHMS IN THIS CHAPTER LOOK LIKE THE FOLLOWING:

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

$t^{(k)}$: STEP SIZE or STEP LENGTH

$\Delta x^{(k)} \in \mathbb{R}^n$, SEARCH DIRECTION

ALL METHODS ARE DESCENT METHODS, IE

$$f(x^{(k+1)}) < f(x^{(k)}) \quad \text{(UNLESS } x^{(k)} \text{ IS OPTIMAL)}$$

BY CONVEXITY WE HAVE $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$,

IE WE MOVE TOWARDS A DESCENT DIRECTION

GENERAL DESCENT METHOD

GIVEN A STARTING POINT $x \in \text{dom } f$

REPEAT

1) FIND Δx
2) LINE SEARCH: CHOOSE $t > 0$
3) UPDATE $x := x + t\Delta x$

UNTIL STOPPING CRITERION (COMMON STOPPING CRITERION: $\|\nabla f(x)\|_2 \leq \eta$)

EXACT LINE SEARCH:

$$t = \operatorname{argmin}_{s \geq 0} f(x + s\Delta x)$$

ONLY MAKES SENSE IF $f(x)$ IS EASY/FAST TO COMPUTE

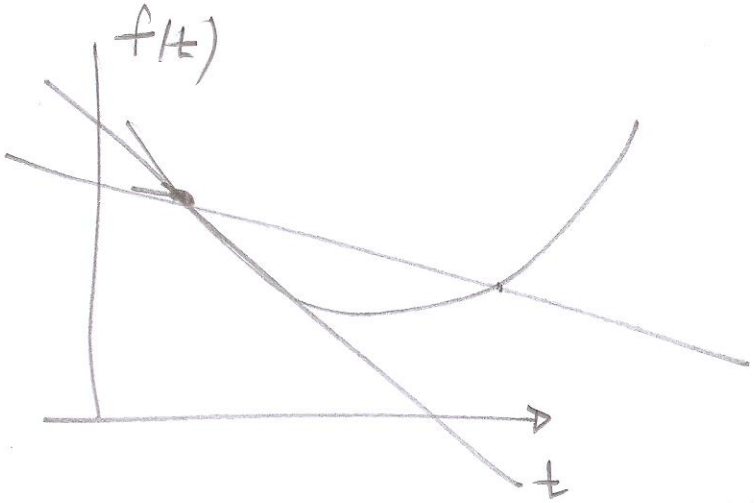
BACKTRACKING LINE SEARCH:

GIVEN Δx FOR f AT $x \in \text{dom } f$, $\alpha \in (0, 0.5)$, $\beta \in (0,1)$

$t := 1$

WHILE $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $t := \beta t$

GEOMETRIC INTERPRETATION:



$$f(x+t\Delta x) \simeq$$
$$f(x) + t\nabla f_{(x)}^T \Delta x$$
$$< f(x) + \alpha t \nabla f_{(x)}^T \Delta x$$

THEREFORE, WE START AT DISTANCE Δx, AND MOVE CLOSER UNTIL WE GET A PERCENTAGE OF THE IMPROVEMENT WE DESIRE.

9.3 GRADIENT DESCENT METHOD

IT IS THE GENERIC DESCENT METHOD, WITH

$$\boxed{\Delta x := -\nabla f_{(x)}}$$

CONVERGENCE ANALYSIS, EXACT LINE SEARCH

(I.E., HOW FAST WE FIND MINIMUM)

WE ASSUME: $mI \leq \nabla^2 f_{(x)} \leq MI \quad \forall x \in S$

WE DEFINE $\tilde{f}(t) = f(x - t\nabla f_{(x)})$

(9.13)
$y = x - t\nabla f_{(x)}$ $\Rightarrow$

$$\tilde{f}(t) \leq f(x) - t\|\nabla f_{(x)}\|_2^2 + \frac{Mt^2}{2}\|\nabla f_{(x)}\|_2^2 \quad (9.17)$$

AT FIRST, LINEAR DECREASE, EVENTUALLY QUADRATIC DECREASE

MINIMIZE THE RIGHT HAND SIDE:

$$-\|\nabla f(x)\|_2^2 + \frac{M t^2}{2}\|\nabla f(x)\|_2^2 \Leftrightarrow t = \frac{1}{M}$$

AS WE STOP AT THE MINIMUM, WE HAVE

$$f(x^{(k+1)}) \le f(x) - \frac{1}{M}\|\nabla f(x)\|_2^2 + \frac{1}{2M}\|\nabla f(x)\|_2^2$$

$$\Rightarrow f(x^{(k+1)}) - p^* \le (f(x) - p^*) - \frac{1}{2M}\|\nabla f(x)\|_2^2$$

HOWEVER, WE KNOW THAT $\|\nabla f(x)\|_2^2 \ge 2m(f(x) - p^*)$
(COMES FROM (9.9))

$$\Rightarrow f(x^{(k+1)}) - p^* \le \left(1 - \frac{m}{M}\right)(f(x) - p^*) \Rightarrow$$

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

$$c = 1 - \frac{m}{M} < 1$$

BETTER

NO RATE OF CONVERGENCE IS GEOMETRIC OR, AS WE SAW IN CONVERGENCE ANALYSIS, LINEAR

THEREFORE, IF WE WANT $f(x^{(k)}) - p^* \le \epsilon$, WE NEED

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

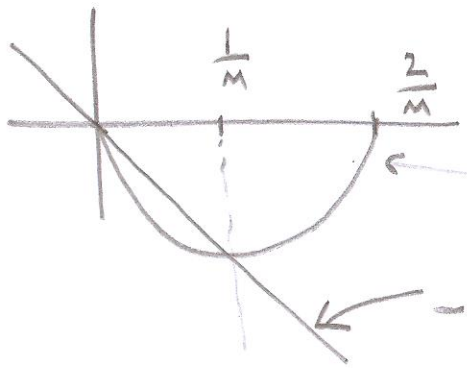
$$c^k \le \frac{\epsilon}{f(x^{(0)}) - p^*} \Leftrightarrow k \log c \le \log\left[\frac{\epsilon}{f(x^{(0)}) - p^*}\right]$$

$$\Leftrightarrow k \ge \frac{\log\left[\frac{\epsilon}{f(x^{(0)}) - p^*}\right]}{\log c} \Leftrightarrow k \simeq \frac{\log\left[\frac{f(x^{(0)}) - p^*}{\epsilon}\right]}{\log\left(\frac{1}{c}\right)}$$

THEREFORE, WE WANT $\frac{m}{M}$ AS LARGE AS POSSIBLE!

ANALYSIS FOR BACKTRACKING LINE SEARCH

We show that backtracking condition is satisfied for $0 < t \le \frac{1}{M}$. This will guarantee that we do not stop too close to the origin $t = 0$.



$\Rightarrow$ $0 \le t \le \frac{1}{M}$

$\Rightarrow$

$-t + \frac{Mt^2}{2} \le -\frac{t}{2}$

We know that

$$\tilde{f}(t) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$$

(for $0 \le t \le \frac{1}{M}$)

$$\le f(x) - \left(\frac{t}{2}\right)\|\nabla f(x)\|_2^2$$

$$\le f(x) - \alpha t \|\nabla f(x)\|_2^2$$

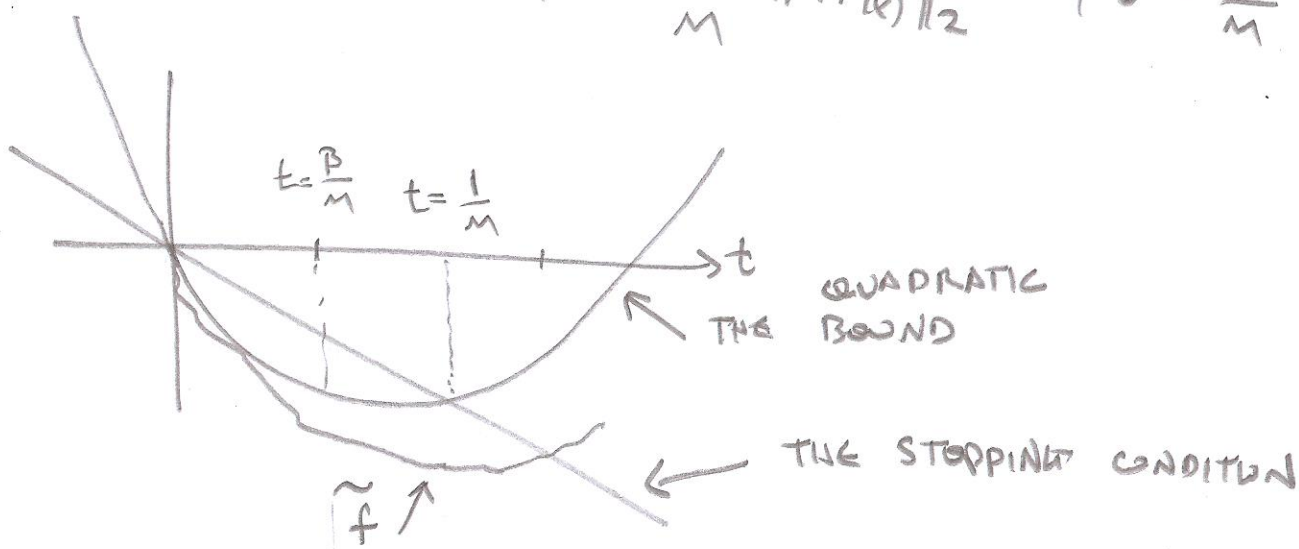
(we assume $\alpha < \frac{1}{2}$)

So the backtracking is guaranteed to have stopped either for $t = 1$, (if $1 \le \frac{1}{M}$) or some (if $1 > \frac{1}{M}$) $t \ge \frac{\beta}{M}$, so we have

$$f(x^+) \le f(x) - \alpha \|\nabla f(x)\|_2^2 \quad (t = 1)$$

or

$$f(x^+) \le f(x) - \frac{\beta\alpha}{M}\|\nabla f(x)\|_2^2 \quad \left(t = \frac{\beta}{M}\right)$$



so $$f(x^+) \le f(x) - \min\left\{\alpha, \frac{\beta\alpha}{M}\right\} \|\nabla f(x)\|_2^2$$

$$\Rightarrow \quad f(x^+) - p^* \le f(x) - p^* - \min\left\{\alpha, \frac{\beta\alpha}{M}\right\} \cdot \|\nabla f(x)\|_2^2$$

$$\downarrow \quad \geq 2m\,(f(x) - p^*)$$

$$\Rightarrow \quad f(x^{(k+1)}) - p^* \le (f(x^{(k)}) - p^*)\left[1 - 2m\min\left\{\alpha, \frac{\beta\alpha}{M}\right\}\right]$$

$$\| \quad c < 1$$

$$\Rightarrow \quad f(x^{(k+1)}) - p^* \le c^k \left[f(x^{(0)}) - p^*\right],$$

so again we have linear convergence

<u>An example (quadratic)</u>

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$$

$$H = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}, \quad \text{so} \quad m = \min(1, \gamma), \quad M = \max(1, \gamma).$$

If we start from $x^{(0)} = (\gamma, 1)$, we can derive closed formula expressions:

$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k$$

$$f(x^{(k)}) = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)})$$

THEREFORE, CONVERGENCE IS INDEED LINEAR (EXCEPT WHEN $\gamma = 1$, IN WHICH CASE WE FIND OPTIMUM IN FIRST TRY.

LET US COMPARE CONVERGENCE WITH PREVIOUS BOUND. (OF THIS EXAMPLE)

WE KNOW THAT, IN GENERAL CASE,

$$f(x^{(k)}) - p^* \leq c^k \left(f(x^{(0)}) - p^* \right),$$

WHERE $\boxed{c = 1 - \frac{m}{M}}$, $mI \leq \nabla^2 f(x) \leq MI$

IN THIS CASE,

$$f(x^{(k)}) - p^* \leq q^k \left(f(x^{(0)}) - p^* \right)$$

WHERE $q = \left(\frac{\gamma - 1}{\gamma + 1} \right)^2$

IF $\gamma > 1$, $\gamma = M$, $L = m \Rightarrow$ $q = \left(\frac{M - m}{M + m} \right)^2 \Rightarrow$

$$q = \left(\frac{1 - \frac{m}{M}}{1 + \frac{m}{M}} \right)^2$$

IF $\gamma < L$, $\gamma = m$, $M = L$, $\Rightarrow$ $q = \left(\frac{m - M}{m + M} \right)^2 = \left(\frac{\frac{m}{M} - 1}{1 + \frac{m}{M}} \right)^2$

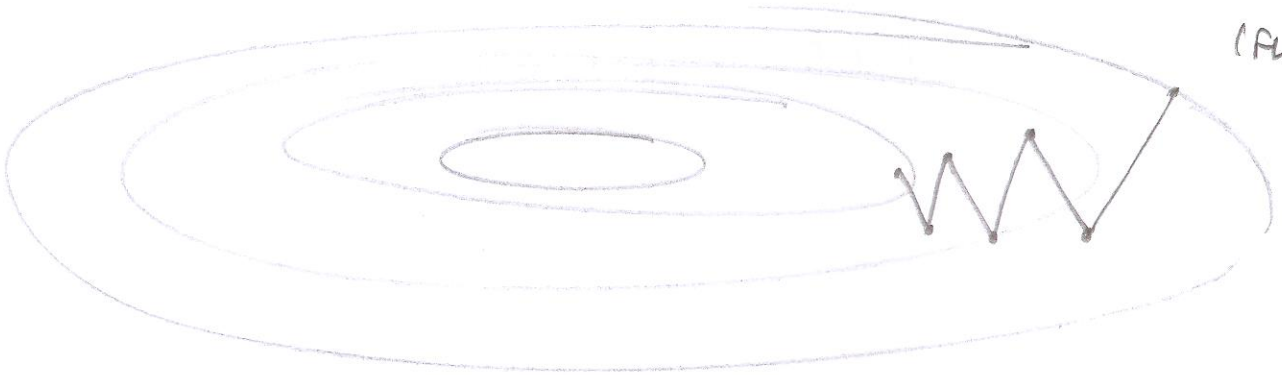
SO, ALWAYS,

$$\boxed{q = \left(\frac{1 - \frac{m}{M}}{1 + \frac{m}{M}} \right)^2}$$

WHICH IS FASTER. IN FACT, FOR SMALL $\frac{m}{M}$, IT IS 4 TIMES FASTER, BECAUSE $q \simeq \left(1 - \frac{m}{M}\right)^2 \left(1 - \frac{m}{M}\right)^2$

USING THE APPROXIMATE $\frac{1}{1+\sigma} \simeq 1-\sigma$, OR $\simeq 0$.

SO MISTAKE IN ANALYSIS IS ONLY BY A FACTOR OF 4.



(FIGURE 9.2)

THE ABOVE SHOWS WHY POOR CONDITION NUMBERS REALLY HURT US.

SOME GENERAL REMARKS:

1) CONVERGENCE IS LINEAR
2) w, β NOT SO CRUCIAL
3) WHEN CONDITION NUMBER IS HIGH, IE, $m \ll M$, THE METHOD IS USELESS.

9.4 STEEPEST DESCENT METHODS

$$f(x+v) \simeq \hat{f}(x+v) = f(x) + \nabla f(x)^T v$$

↑ DIRECTIONAL DERIVATIVE IN DIRECTION v

DEFINITION:
LET $\|\cdot\|$ BE ANY NORM. WE SELECT THE

NORMALIZED STEEPEST DESCENT DIRECTION

$$\Delta x_{nsd} = \operatorname{argmin}\{\nabla f(x)^T v \mid \|v\| = 1\}$$
$$= \operatorname{argmin}\{\nabla f(x)^T v \mid \|v\| \leq 1\}$$

DEFINITION: THE STEEPEST DESCENT METHOD IS LIKE THE GRADIENT DESCENT METHOD, BUT SELECTS AS THE DESCENT DIRECTION

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

Δx_{sd}: UNNORMALIZED STEEPEST DESCENT

$\|\cdot\|_*$: DUAL NORM, ENSURES THAT OF THE NORM USED, STEP IS ADEQUATELY AGGRESSIVE

VARIOUS CHOICES EXIST FOR THE CHOICE OF THE USED NORM:

(1) NORM IS EUCLIDEAN: THEN STEEPEST DESCENT = GRADIENT DESCENT

(2) NORM IS QUADRATIC NORM:

$$\|z\|_P = (z^T P z)^{1/2} = \|P^{\frac{1}{2}} z\|_2$$

WHERE $P \in S^n_{++}$. NOTE: AS $P \in S^n_{++}$, WE CAN ALWAYS WRITE

$$P = Q^T \Lambda Q$$

(Q: ORTHONORMAL) $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $\lambda_i > 0$,

SO $$P^{\frac{1}{2}} \triangleq Q^T \mathrm{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_n^{1/2})$$

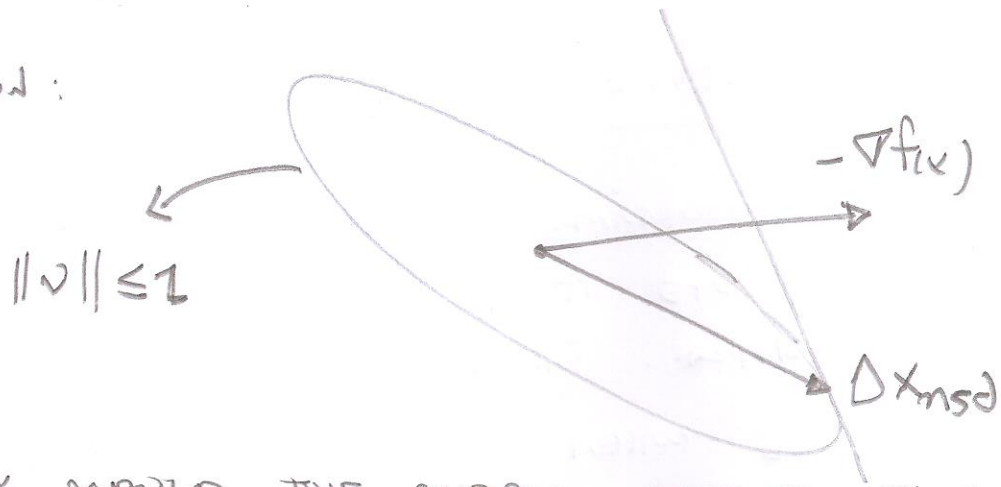
IN THIS CASE, WE CAN SHOW:

$$\Delta x_{nsd} = -\left(\nabla f_{(x)}^T P^{-1} \nabla f_{(x)}\right)^{-\frac{1}{2}} P^{-1} \nabla f_{(x)},$$

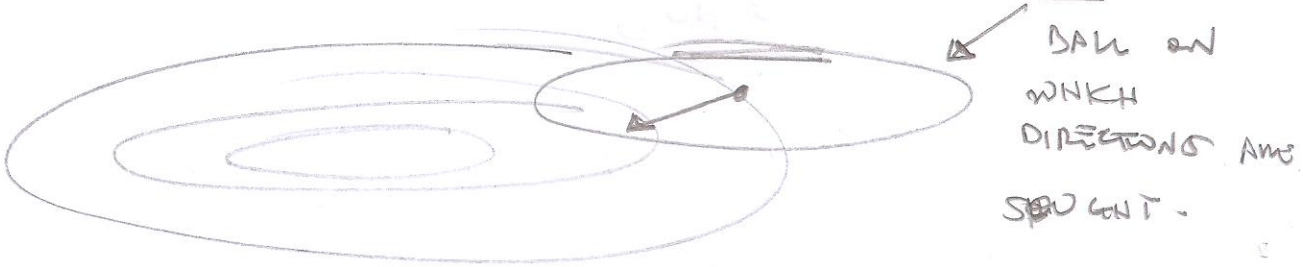
$$\Delta x_{sd} = -P^{-1} \nabla f_{(x)},$$

BY SOLVING A CONSTRAINT OPTIMIZATION PROBLEM

INTUITION:



SO, WHY WOULD THIS WORK BETTER? REMEMBER CASE OF ILL-CONDITIONING:



OF COURSE, WE COULD DO MUCH WORSE!
IDEALLY, WE WANT TO MATCH WITH OUR UNIT BALL OUR HESSIAN AT THE OPTIMUM.

(3) NORM IS ℓ_1

$\|z\|_1 = |z_1| + |z_2| + \ldots + |z_n|$ IN THIS CASE,

$$\Delta x_{nsd} = \text{arg min}\left\{\nabla f_{(x)}^T v \mid \|v\|_1 \leq 1\right\}.$$

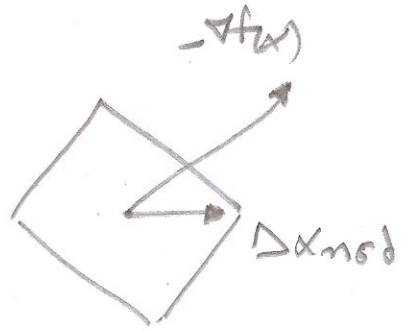
ONE OPTIMUM IS: $\Delta x_{nsd} = -\text{sign}\left(\frac{\partial f(x)}{\partial x_i}\right) e_i$

(OF THE POSSIBLE MANY)

DUAL OF ℓ_1 NORM

$$\Rightarrow \quad \Delta x_{sd} = \Delta x_{nsd} \, \|\nabla f(x)\|_\infty = -\frac{\partial f(x)}{\partial x_i} e_i$$

INTUITION:



WE CAN ALWAYS PICK A DIRECTION IN A STANDARD BASIS VECTOR.

$\Rightarrow$ COORDINATE-DESCENT!

ALGORITHM: FIND COORDINATE ON WHICH DESCENT IS THE STEEPEST, AND USE IT TO DO EXACT LINE SEARCH OR BACKTRACK SEARCH. THEN REPEAT.

THIS IS VERY USEFUL WHEN SOLVING MINIMIZATION PROBLEM ACROSS ONE COORDINATE IS EASY.

9.5 NEWTON'S METHOD

DEFINITION: NEWTON STEP: $\Delta x_{nt} = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$.

IT IS CERTAINLY A DESCENT DIRECTION, BECAUSE

$$\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \left[\nabla^2 f(x)\right]^{-1} \nabla f(x) < 0,$$

SINCE $\left[\nabla^2 f(x)\right]^{-1}$ IS POSITIVE DEFINITE. THIS IS THE STEP WE WE WILL USE FOR OUR GRADIENT DESCENT ALGORITHM. WHY?

1) IT IS THE MINIMIZER OF THE SECOND ORDER APPROXIMATION

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

SO LET'S TRY TO MINIMIZE THAT!

$$\nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0 \Rightarrow v = -\left[\nabla^2 f(x)\right]^{-1} \nabla f(x)$$

2) IT IS THE STEEPEST DESCENT DIRECTION USING THE HESSIAN NORM (WHICH WE KNOW IS A VERY GOOD CHOICE FOR THE S.D. NORM). INDEED:

FROM STEEPEST DESCENT THEORY, IF WE USE $\|z\|_P = (z^T P z)^{\frac{1}{2}} = \|P^{\frac{1}{2}} z\|_2$, THEN $\Delta x_{SD} = -P^{-1}\nabla f(x)$

3) IT IS THE SOLUTION OF THE LINEARIZED OPTIMALITY CONDITION

OBSERVE THAT

$$\nabla f(x+v) \cong \nabla f(x) + \nabla^2 f(x)\, v$$

INDEED, WE $F(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}$. THEN

$$F(x+v) \cong F(x) + \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial x_m}{\partial x_m} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix},$$

AND SUBSTITUTING $f(x)$ FOR ∇f, THE ABOVE RESULT FOLLOWS

SO MINIMIZING THE ABOVE, THE RESULT FOLLOWS.

IN OTHER WORDS, THIS IS AN ESTIMATION OF WHERE THE OPTIMALITY CONDITIONS HOLD

A BASIC PROPERTY OF THE NEWTON INCREMENT: IT IS INVARIANT TO AFFINE TRANSFORMATIONS. LET $T \in \mathbb{R}^{n \times n}$, $\det T \neq 0$.

LET $\bar{f}(y) = f(Ty) \Rightarrow \begin{cases} \nabla \bar{f}(y) = T^T \nabla f(x) \\ \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T \end{cases}$

$\boxed{x = Ty}$

NEWTON STEP FOR $\bar{f}$ AT y IS:

$$\Delta y_{nt} = -\left(T^T \nabla^2 f(x) T\right)^{-1} \left(T^T \nabla f(x)\right)$$

$$= -T^{-1} \left[\nabla^2 f(x)\right]^{-1} \nabla f(x) = T^{-1} \Delta x_{nt} \Rightarrow$$

$$\boxed{\Delta x_{nt} = T \Delta y_{nt}}$$

$$\Rightarrow \boxed{x + \Delta x_{nt} = T\left(y + \Delta y_{nt}\right)}$$

THIS IMPLIES THAT NEWTON'S METHOD, TO BE INTRODUCED BELOW, IS NOT AFFECTED BY COORDINATE TRANSFORMATIONS.

ALGORITHM: NEWTON'S METHOD

GIVEN x STARTING POINT AND TOLERANCE $\epsilon > 0$.

REPEAT

1) COMPUTE { NEWTON STEP $\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x)$

NEWTON DECREMENT $\lambda^2 := \nabla f(x)^T \nabla^2 f(x) \nabla f(x)$

2) STOPPING CRITERION: QUIT IF $\frac{\lambda^2}{2} \leq \epsilon$

3) LINE SEARCH: CHOOSE STEP SIZE t BY BACKTRACKING

4) UPDATE $x := x + t \Delta x_{nt}$

COMMENTS:

1) IF THERE IS NO BACKTRACK, AND $t=1$ ALWAYS, THEN WE CALL THE METHOD THE PURE NEWTON'S METHOD.

2) REGARDING THE STOPPING CRITERION:

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x+\Delta x_{nt}) =$$

$$-\nabla f(x)^T \left[-\nabla^2 f(x)^{-1}\right] \nabla f(x)$$

$$- \frac{1}{2} \nabla f(x)^T \left[-\nabla^2 f(x)^{-1}\right]^T \nabla^2 f(x) \left[-\nabla^2 f(x)^{-1} \nabla f(x)\right]$$

$$= +\frac{1}{2} \nabla f(x)^T \left[\nabla^2 f(x)^{-1}\right] \nabla f(x) = \frac{\lambda^2}{2}$$

THE QUADRATIC APPROXIMATION of f

SO $\frac{\lambda^2}{2}$ IS AN ESTIMATE OF OUR ERROR

3) ONCE WE ARE CLOSE ENOUGH TO THE SOLUTION, THE CONVERGENCE IS QUADRATIC!

ADVANTAGES OF NEWTON'S METHOD

1) RAPID CONVERGENCE
2) AFFINE INVARIANT
3) SCALES WELL
4) DOES NOT DEPEND ON CHOICE OF COORDINATES

DISADVANTAGE: YOU NEED TO FIND A MATRIX (THE HESSIAN) AND INVERT IT!

⇒ QUASI-NEWTON METHODS HAVE BEEN DEVELOPED, WHICH AVOID HEAVY COMPUTATION