2. Line search methods

2.1 Principle of the LS method

2.1.1 ⊛ $x^{k+1} = x^k + \alpha_k p^k$: iteration

$p^k = -B_k^{-1} \nabla f(x^k)$: search direction

Here B_k is symmetric and nonsingular, and p.s.d in virtually all approaches

~~Particular cases:~~

~~$B_k = I$: "steepest descent"~~

~~$B_k = \nabla^2 f(x^k)$: Newton's method~~

⊛ $\boxed{\text{The descent condition}}$

Use Taylor's theorem:

$$f(x^k + \alpha p^k) = f(x^k) + \alpha p^T \nabla f(x^k) + \frac{1}{2} \alpha^2 p^T \nabla^2 f(x^k + t p^k) p$$

for $t \in (0, \alpha)$

Mihai Anitescu Notes

2.1.2

$$\frac{d}{d\alpha} f(x^k + \alpha p^k) = \nabla f(x^k)^T p^k$$

⊛ So: $\nabla f(x^k)^T p^k < 0 \Rightarrow p^k$ is a descent direction

$\exists\, \varepsilon$ s.t $f(x^k + \varepsilon p^k) < f(x^k)$

⊛ In particular:

$$\min \nabla f(x^k)^T p^k \quad \text{s.t } p^{k^T} p^k = 1 \Rightarrow p^k = \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

$\Downarrow$ "steepest descent"

⊛ Since $\hat{p}^k = t p^k$ induces same search direction with different scaling, we might as well take $p^k = -\nabla f(x^k)$

$\Leftrightarrow B^k = I$.

⊛ IDEA: use $f(x)$ as a Lyapunov function $\boxed{\dot{x} = -\nabla f(x)}$ to keep going down hill

Mihai Anitescu Notes

2.1.3. ⊗ In conclusion:

$B_k = I \Rightarrow$ steepest descent

$B_k = \nabla^2 f(x^k) \Rightarrow$ Newton's method but may not be eligible as a descent direction

$B_k = \nabla^2 f(x^k) + \gamma^k I \Rightarrow$ modified Newton (sometimes quasi-Newton)

Typically $B_k \succeq 0$ and $\gamma^k \to 0$ as $k \to \infty$

Mihai Anitescu Notes

2. Line search

2.2. Computing step length

2.2.1 Ideally

$$\min_{\alpha \geq 0} f(x^k + \alpha p^k)$$

"Exact line search" => not practical

- One solution: Inexact line search
- But how inexact?

Is it sufficient to have

$$f(x^k + \alpha p^k) < f(x^k)$$

No. E.g. Book: $f(x^k) = 5/k$

$$f(x^*) = -1$$

Mihai Anitescu Notes

2 Line Search

2.3 Wolfe Conditions

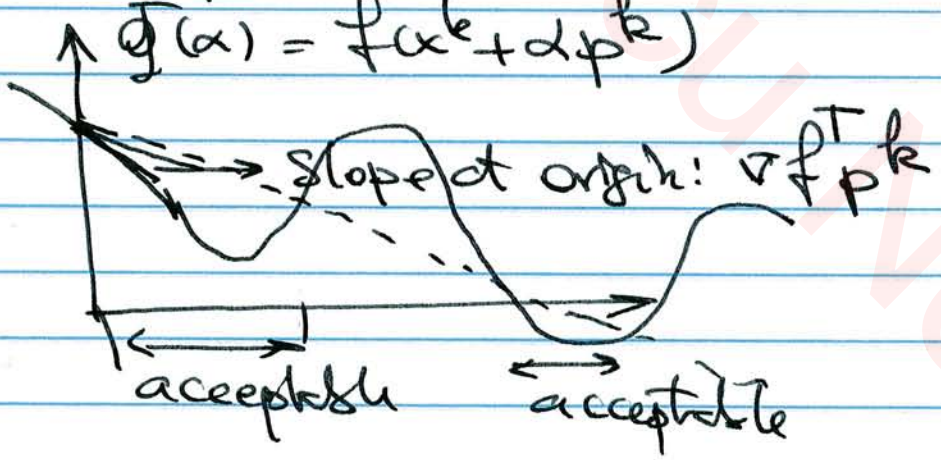
2.3.1 We need <u>sufficient decrease</u>

$$f(x^k + \alpha p^k) \le f(x^k) + c_1 \alpha \nabla f(x^k)^T p^k$$

$c_1 \in (0,1)$

Typically $c_1 \sim 10^{-2} - 10^{-4}$ if use a Newton method

Graphical Representation



Problem: any small step can still be acceptable.

Mihai Anitescu Notes

2.3.2 To rule out very small steps, we do one of the 2:

1. Choose "~~largest~~ smallest" $k \geq 0$ such that ($\rho < 1$)

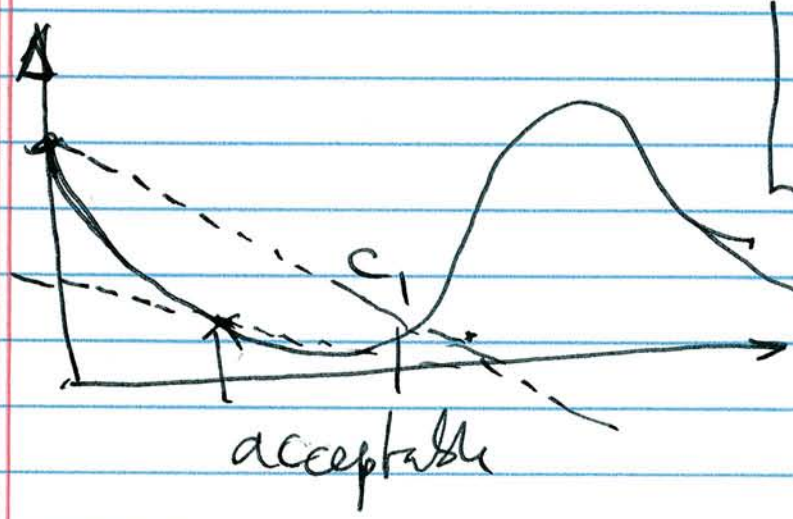
$$f(x^k + \rho^k \bar{\alpha} p^k) \leq f(x^k) + c_1 \bar{\alpha} \rho^k \nabla f(x^k)^T p^k$$

Armijo Rule

2. Curvature Condition

$$\nabla f(x^k + \alpha_k p^k)^T p^k \geq c_2 \nabla f(x^k)^T p^k$$

Negative



Slope less than a fraction at origin

Milan Antescu Notes

2.3.3 (*) Strong Wolfe Condition

$$f(x^k+\alpha p^k) \le f(x^k) + c_1 \alpha \nabla f(x^k)^T p^k$$

"sufficiently underneath level"

$$|\nabla f(x^k+\alpha p^k)^T p^k| \le c_2 \cdot |\nabla f(x^k)^T p^k|$$

$$\parallel$$

$$\frac{d}{d\alpha} f(x^k+\alpha p^k)$$

"suff close to minimum

(*) How large should c_1 be?

Ideally; global min should be eligible, should satisfy strong Wolfe conditions

$\cancel{f(x^*)}$

$\boxed{\text{Curvature Condition - Works}}$

Mihai Anitescu Notes

2.3.4 How about ~~global~~ suf decrease?

$$f(x^*) \le f(x^k) + c_1 \cdot \frac{\|x^* - x^k\|}{\|p^k\|} \cdot \nabla f(x^k)^T p^k$$

f: quadratic $\Rightarrow$

$$\phi(\alpha^*) \le \phi(0) + c_1 \cdot \alpha^* \phi'(0)$$

$$0 \le +\alpha^* - c_1 \cdot \alpha^* \cdot 2\alpha^*$$

$$\Rightarrow \boxed{c_1 \le 1/2}$$

Lemma 3.1. $f: \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let p^k be a descent direction at x^k and assume f is bdd below along $\{x^k + \alpha p^k, \alpha > 0\}$.

Then: $0 < c_1 < c_2 < 1$ $\exists$ interval of α satisfying Wolfe.

Minai Antiescu Notes

2.3.5 Proof : One d.

$$\begin{cases} \phi(\alpha) = f(x^k + \alpha p^k) \\ \phi(\alpha) \le c_1 \alpha \phi'(0) + \phi(0) \\ |\phi'(\alpha)| \le c_2 |\phi'(0)| \end{cases}$$

Descent direction $\Rightarrow \phi'(0) < 0$

$$\alpha' = \min_{\alpha > 0} \{ \alpha > 0 \mid \phi(\alpha) = \phi(0) + c_1 \alpha \phi'(0) \}$$

$\Rightarrow \alpha' > 0$ [smallest $\alpha > 0$]

∞

MVT $\exists \ \alpha'' \in (0, \alpha')$

s.t. $\phi'(\alpha'') = \dfrac{\phi(\alpha) - \phi(0)}{\alpha} = c_1 \phi'(0)$

$|\phi'(\alpha'')| = |c_1 \phi'(0)| < c_2 |\phi'(0)|$

Mihai Anitescu Notes

2.3.6 But g' is cts $\Rightarrow$ $\exists$ interval around g^4 such that Wolfe holds.

Mihai Anitescu Notes

2. Line Search

2.4. Backtracking

2.4.1

Choose $\bar{\alpha} > 0$; $\rho \in (0,1)$, $c \in (0,1)$

$\alpha = \bar{\alpha}$

repeat until $f(x^k + \alpha p^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T p^k$

$\alpha \leftarrow \rho \alpha$

end

$\alpha^k = \alpha$

If $B_k = \nabla^2 f(x^k)$ (Newton) start with $\bar{\alpha} = 1$

$c_1 = 10^{-3} - 10^{-1}$ (I used 0.3 once)

$\rho = 0.2 - 0.8$

You will have to implement it :-)

Miguel Antelecu Notes

2. Line Search Methods

2.5 Convergence of L.S.M.

2.5.1

Theorem 3.2 (Zoutendijk)

0 (Wolfe condition) : can be replaced by S BB

Assume 1. p^k is a descent direction

2. $f \in C^1$ in $\mathcal{N}(\{x \mid f(x) \le f(x^0)\})$

3. ∇f is Lip on $\mathcal{N}$ with L

4. f is bounded below in $\mathbb{R}^n$.

Then $$\sum_{k \ge 0} \cos^2\theta_k \|\nabla f(x^k)\|^2 < \infty$$

Here. $$\cos\theta_k = \frac{p_k^{kT} \nabla f(x^k)}{\|p_k^k\| \cdot \|\nabla f(x^k)\|}$$

Proof

(Wolfe L)

$$\nabla f(x^{k+1})^T p^k - \nabla f(x^k)^T p^k \ge (c_2 - 1) \nabla f(x^k)^T p^k$$

(Lipshitz)

$$\nabla f(x^{k+1})^T p^k - \nabla f(x^k)^T p^k \le L \cdot \alpha_k \cdot \|p^k\|^2$$

$$\Rightarrow L \alpha_k \cdot \|p^k\| \ge (c_2 - 1) \nabla f(x^k)^T p^k$$

$$\Rightarrow \alpha_k \ge \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^T p^k}{\|p^k\|^2}$$

2.5.2

$$f(x^{k+1}) \leq f(x^k) + c_1 \alpha^k \cdot \nabla f(x^k)^T p^k$$

But $\nabla f(x^k)^T p^k = \cos(\theta_k) \cdot \|\nabla f(x^k)\| \cdot \|p^k\|$

Also $\alpha^k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^T p^k}{\|p^k\|^2} =$

$$\Rightarrow \frac{c_2 - 1}{L} \cdot \frac{\cos(\theta_k) \cdot \|\nabla f(x^k)\|}{\|p^k\|}$$

$$\Rightarrow f(x^{k+1}) \leq f(x^k) + c_1 \alpha_k \nabla f(x^k)^T p^k$$

$$\leq f(x^k) + c_1 \frac{c_2 - 1}{L} \|\nabla f(x^k)\|^2 \cdot \cos(\theta_k)^2$$

$$\Rightarrow f(x^K) \leq f(x^0) - \left(\frac{1 - c_2}{L} c_1\right) \cdot \sum_{k=0}^{N} \left[\cos(\theta_k)\right]^2 \|\nabla f(x^k)\|^2$$

$$\Rightarrow \underbrace{c_3} \cdot \sum_{k=0}^{N} \left[\cos(\theta_k)\right]^2 \cdot \|\nabla f(x^k)\|^2 \leq f(x^0) - f(x^N)$$

Bounded Below $< \infty$

Q.E.D.

2.5.3 Implications

A) Global convergence (weak) of line search methods

$$\cos(\theta_k)^2 \cdot \|\nabla f(x_k)\|^2 \to 0$$

Assume $|\cos(\theta_k)| \geq \delta$: descent is of

$\Rightarrow$ sufficient quality.

Then $\nabla f(x_k) \to 0$ (weak global convergence)

B) Armijo Rule.

Note that the only purpose of Wolfe B is to ensure step is bounded below.

{ You can show $\boxed{\text{Lipshitz} + |\cos(\theta_k)| \geq \delta}$

$\Rightarrow \alpha^k \geq c > 0$.

Then convergence also applies by a small modification.

2.5.4 Further implications.

C) For line search, it is impossible to guarant
global convergence to SOSC point.

Think: start at stationay point.

So I must introduce some curvature information.

D) How to guarantee $|\cos\theta_k| \geq \delta > 0$?

Assume $\|B_k\| \cdot \|B_k^{-1}\| \leq M$ $\Rightarrow$ (homework)

↳ aka "condition number"

$$\cos(\theta_k) = \frac{\nabla f(x^k)^T p^k}{\|\nabla f(x^k)^T\| \cdot \|p^k\|}$$

$p^k = -B_k^{-1} \nabla f(x^k)^T \Rightarrow$ Use eigs. $\Leftarrow$ $\dfrac{\nabla f(x^k)^T B_k^{-1} \nabla f(x^k)}{\|\nabla f(x^k)^T\| \|B_k^{-1} \nabla f(x^k)\|}$

2. Line search methods

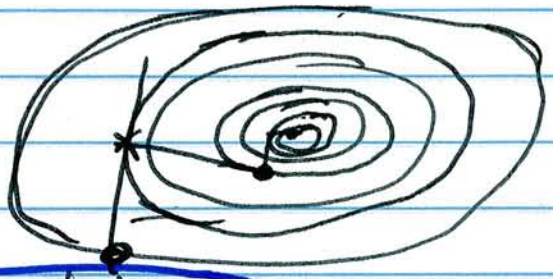
2.6. Steepest descent method

Exclude if out of time

(k) $p^k = -B_k \nabla f(x^k)$; choose $B_k = I$.

(k) On a quadratic ; graphical interpretation

The zig-zag behavior is typical, and it is what makes it slow.



Exclude if out of time

(k) Theorem. When steepest descent with exact line searches is applied to the strongly convex quadratic function

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

we have

$$\|x\|_Q = \sqrt{x^T Q x}$$

$$\|x^{k+1} - x^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 \|x^k - x^*\|_Q^2$$

2.6.2 Theorem 3.4. Assume that

(1) $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and (2) that the iterates generated by the steepest descent with exact line search converges to a point that satisfies SOSC ($\nabla^2 f(x^*) \succ 0$). Let r be a scalar $r \in \left(\frac{d_n - d_1}{d_n + d_1}, 1\right)$; $[d_1, d_n] \supset \sigma(\nabla^2 f(x^*))$.

Then $\forall$ k suf. large, we have

$$\left[f(x^{k+1}) - f(x^*)\right] \leq r^2 \cdot \left[f(x^k) - f(x^*)\right]$$

Discussion. Cannot expect it to be better

$$c(\nabla f^2(x^*)) = \left[\frac{d_n}{d_1} = 1000\right] \to \text{Not BAD} \Rightarrow \frac{d_n - d_1}{d_n + d_1} =$$

$$\frac{1 - \frac{d_1}{d_n}}{1 + \frac{d_1}{d_n}} \simeq 1 - 2\frac{d_1}{d_n} \Rightarrow \left(1 - \frac{2d_1}{d_n}\right)^2 \simeq 1 - \frac{4d_1}{d_n}$$

2.6.3

$$f(x^{k+1}) - f(x^*) \le 0.996\,[f(x^k) - f(x^*)]$$

$$(0.996)^k \le 0.001 \Rightarrow k \simeq \frac{-3 \cdot \ln 10}{0.004} \simeq 1000$$

$$\ln(0.996) \simeq 0.004$$

iterations!!

Of course, upper bound, but can be shown it is attainable

What is the problem? Steepest descent is NOT scaled. $\Rightarrow$ Newton

Mihai Anitescu Notes

2. Line search methods

2.7. "Almost" Newton methods

(A) $p_k = -B_k^{-1} \nabla f(x^k)$. $k = 0, 1, 2, \ldots$

(A) $B_k = \nabla^2 f(x)$ is hard to get and maybe not a descent direction.

Idea: Modify it. Maybe $B_k = \nu_k I + \nabla^2 f(x^k)$, $\nu_k \to 0$ but $\nu_k + \nabla^2 f(x^k) \geq \varepsilon I$?

[E.g. $\nu_k = 2 \cdot \min\{\min \lambda(\nabla^2 f), 0\}$]

But: Do I still obtain fast convergence (we saw linear for S.D. is slow).

Theorem 3.7 Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is C^2. $x^{k+1} = x^k + p^k$ ($\alpha^k = 1$ satisfies Wolfe or Armijo, e.g.). Assume $x^k \to x^*$ s.t. x^* satisfies SOSC. Then x^k converges superlinearly $\Leftrightarrow$

2.7.2 $$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 f(x^k))^T p^k\|}{\|p^k\|} = 0 \quad (DM)$$

Def

Discussion Interestingly, it is sufficient for B_k to look like $\nabla^2 f(x^k)$ only along the search direction p^k. Methods such quasi-Newton satisfy this.

Proof (DM)

Assume near convergence so it exists

a) $(DM) \Leftrightarrow p^k - p^{k,N} = o(\|p^k\|)$

Recall i) $p^{k,N} = -\nabla^2 f(x^k)^{-1} \cdot p^k$

ii) $y = o(x) \overset{\Delta}{\Leftrightarrow} \lim_{\substack{x\to 0 \\ x>0}} \frac{y}{\|x\|} = 0$

2.7.3

their definition

$$p^k - p^{k,N} = +\nabla^2 f(x^k)^{-1} \nabla f(x^k) + B_k p^k$$

$$= \nabla^2 f(x^k)^{-1} \cdot \left[-B_k p^k + \nabla^2 f(x^k) p^k \right]$$

$$= \nabla^2 f(x^k)^{-1} \cdot \underbrace{\left[\nabla^2 f(x^k) - B_k \right] p^k}_{r^k}$$

Asymptotically bounded above and below

$\|B_k\| \leq \|\nabla^2 f(x^k)\|$

$\|B_k^N\| \leq 2\sigma_n$; $\|B_k^{N,M}\| \leq 2 \cdot \frac{1}{\sigma_1}$

$$\frac{1}{2\sigma_n} \|r^k\| \leq \|p^k - p^{k,N}\| \leq \frac{2}{\sigma_1} \|r^k\|$$

So $\|p^k - p^{k,N}\| = o(p^k) \iff \|r^k\| = o(p^k)$

2.7.4 (3.36) $\Rightarrow$ S.L. $\rightarrow$ triangle

(*) $\|x^k + p^k - x^*\| \le \|x^k + p^{k,N} - x^*\| +$

$+ \|p^{k,N} - p^k\| \le O(\|x^k - x^*\|^2) + o(\|p^k\|)$

$O(\|x^k - x^*\|^2)$ ↑ Newton; $o(\|p^k\|)$ ↑ 3.36

Pout $|\|a\| - \|b\|| \le \|a - b\|$

$\Rightarrow \|p^k\| \le \|x^k + p^k - x^*\| + \|x^k - x^*\|$

Use in (*)

$\Rightarrow \|p^k\| - o(\|p^k\|) \le \|x^k - x^*\| + O(\|x^k - x^*\|^2)$

$\Rightarrow \|p^k\| = O(\|x^k - x^*\|)$

Return to (*):

$$\|x^k + p^k - x^*\| \le o(\|x^k - x^*\|)$$

QED

Mihai Anitescu Notes

$$(x^k + p^k - x^*) - (x^k + p^{k,N} - x^*)$$

2.7.5

$$\|p^k - p^{k,N}\| \le \|x^k + p^k - x^*\| + \|x^k + p^{k,N} - x^*\|$$

$$\le \sigma(\|x^k - x^*\|) \Rightarrow \boxed{\cancel{\sigma(}}$$

Also $\|x^k + p^k - x^*\| \ge \big|\|x^k - x^*\| - \|p^k\|\big|$

$\sigma(\|x^k - x^*\|) \ge$ $\Rightarrow$ $\boxed{\cancel{\|p^k\| = O(\|x^* - x}}$

$$\|x^k - x^*\| - \sigma(\|x^k - x^*\|) \le \|p^k\|$$

$$\Rightarrow \|p^k - p^{k,N}\| = \sigma(\|p^k\|)$$

Q.E.D.

Mihai Anitescu Notes

2. Line Search Methods

2.8. Newton's Method with Hessian modification.

Aim: Need to modify it to ensure that p^k is a descent direction.

$$B^k p^k = -\nabla f(x^k)$$

At iteration k:

- $B^k = LL^T$ where $B^k = \nabla^2 f(x^k) + E^k$ is "sufficiently P.D." → e.g. by Cholesky comply w.o. large pivots
- Solve $B^k p^k = -\nabla f(x^k)$
- $x^{k+1} = x^k + \alpha_k p^k$ by Wolfe or Armijo

Theorem 3.8. $f \in C^2$; and $\{x \mid f(x) \le f(x^0)\}$ is compact. Then if $\|B^k\| \|B^{k,-1}\| \le c \ \forall k$ then $\lim_{k\to\infty} \nabla f(x^k) = 0$.

"Any accumulation point is stationary".

2.8.2 Modifications

1. Cholesky ~~for~~ with added multiple of identity (for A)

Attempt $LL^T = A + \tau_k I$
if succes return L
else adjust k

Adjustment

$\min_i a_{ii} > 0 \Rightarrow \tau_0 = 0$

else ($\beta \sim 10^{-3}$) $\tau_0 = -\min(a_{ii}) + \beta$

if fails: $\tau_{k+1} = \max(2\tau_k, \beta)$

The easiest; I will ask you to implement it

Minal Antesch Notes

2.8.3 Modification:

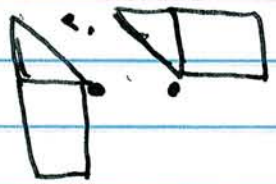
Maybe start with z^k value from previous iteration...

2. Modified Cholesky

Algorithm 3.4: Cholesky (LDL^T form)

for $j = 1, 2 \ldots n$

$$c_{jj} \leftarrow a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2$$



$d_j = c_{jj}$

for $i = j+1, \ldots, n$

$$c_{ij} \leftarrow a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}$$

$$l_{ij} \leftarrow c_{ij} / d_j$$

end

end

Mihai Anitescu Notes

2.8.4 Modification:

$$d_j = \max\left(|c_{jj}|, \left(\frac{\theta_j}{\beta}\right)^2, \delta\right)$$

$$\theta_j = \max_{j<i\leq n} |c_{ij}|$$

But ~~tricks seem~~ it is helpful only if you do pivoting: You do not want to alter unless no choice !!

So should pivot:

$$PAP^T + E = LDL^T$$

Sophisticated, one of better ones, but I do not want to grade it :-)

Minai Antiescu Notes

2. Line Search Methods

2.9. Extensions → exclude wro time

* Interpolation
* Initial Step length non-Newton
* Line Search for Wolfe.

Mihai Anitescu Notes