Numerical Optimisation: Quasi-Newton methods

Marta M. Betcke
m.betcke@ucl.ac.uk,

Kiko Rullan
f.rullan@cs.ucl.ac.uk

Department of Computer Science,
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 7 & 8

Quasi-Newton

- First idea by William C. Davidon in mid 1950, who was frustrated by performance of coordinate descent.
- Quickly picked up by Fletcher and Powell who demonstrated that the new algorithm was much faster and more reliable then existing methods.
- Davidon's original paper was not accepted for publication. More than 30 years later it appeared in the first issue of the SIAM Journal on Optimization in 1991.
- Like steepest gradient, Quasi Newton methods only require the gradient of the objective function at each iterate. Measuring changes in gradient they build a model of the objective function which is good enough to produce superlinear convergence.
- As the Hessian is not required, Quasi-Newton methods can be more efficient than Newton methods which take a long time to evaluate the Hessian and solve for the Newton direction.

Broyden Fletcher Goldfarb Shanno (BFGS)

Quadratic model of the objective function at x_k:

$$m_k(p) = f_k + \nabla f_k^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} B_k p,$$

where $B_k \in \mathbb{R}^{n \times n}$ symmetric positive definite which will be updated during the iteration.

The minimiser of m_k can be written explicitly

$$p_k = -B_k^{-1} \nabla f_k.$$

p_k is used as a search direction and the next iterate becomes

$$x_{k+1} = x_k + \alpha_k p_k.$$

The step length α_k is chosen to satisfy the Wolfe conditions.

The iteration is similar to the line search Newton with the key difference that the Hessian B_k is an approximation.

B_k update

Davidon proposed to update B_k in each iteration instead of computing it anew.

Question: When we computed the new iterate x_{k+1} and construct the new model

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} B_{k+1} p,$$

what requirements should we impose on B_{k+1} based on the knowledge gathered in the last step?

Require: gradient of m_{k+1} should match the gradient of f at the last two iterates x_k, x_{k+1}.

i) At x_{k+1}: $p_{k+1} = 0$,
$\nabla m_{k+1}(0) = \nabla f_{k+1}$ is satisfied automatically.

ii) At $x_k = x_{k+1} - \alpha_k p_k$:

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k.$$

By rearranging ii) we obtain

$$B_{k+1}\alpha_k p_k = \nabla f_{k+1} - \nabla f_k.$$

Define vectors

$$s_k = x_{k+1} - x_k = \alpha_k p_k, \quad y_k = \nabla f_{k+1} - \nabla f_k,$$

ii) becomes the *secant equation*

$$B_{k+1}s_k = y_k.$$

As B_{k+1} is symmetric positive definite, this is only possible if the *curvature condition* holds

$$s_k^{\mathrm{T}} y_k > 0,$$

which can be easily seen multiplying the secant equation by s_k^{T} from the left.

If f is strongly convex $s_k^{\mathrm{T}} y_k > 0$ is satisfied for any x_k, x_{k+1}. However, for nonconvex functions in general this condition will have to be enforced explicitly by imposing restrictions on the line search.

$s_k^{\mathrm{T}} y_k > 0$ is guaranteed if we impose Wolfe or strong Wolfe conditions:

From the 2nd Wolfe condition $s_k^{\mathrm{T}} \nabla f_{k+1} \geq c_2 s_k^{\mathrm{T}} \nabla f_k$, $c_1 < c_2 < 1$ it follows

$$s_k^{\mathrm{T}} y_k \geq (c_2 - 1)\alpha_k p_k^{\mathrm{T}} \nabla f_k > 0,$$

since $c_2 < 1$ and p_k is a descent direction, and the curvature condition holds.

Davidon Flecher Powell (DFP)

When $s_k^{\mathrm{T}} y_k > 0$, the secant equation always has a solution B_{k+1}. In fact the secant equation is heavily underdetermined: a symmetric matrix has $n(n+1)/2$ dofs, secant equation: n conditions, positive definiteness: n inequalities.

Extra conditions to obtain unique solutions: we look for B_{k+1} close to B_k in a certain sense.

DFP update:

$$B_{k+1} = (I - \rho_k y_k s_k^{\mathrm{T}}) B_k (I - \rho_k s_k y_k^{\mathrm{T}}) + \rho_k y_k y_k^{\mathrm{T}} \qquad \text{(DFP B)}$$

with $\rho_k = 1/y_k^{\mathrm{T}} s_k$.

The inverse $H_k = B_k^{-1}$ can be obtained with Sherman-Morrison-Woodbury formula

$$H_{k+1} = H_k - \frac{H_k y_k y_k^{\mathrm{T}} H_k}{y_k^{\mathrm{T}} H_k y_k} + \frac{s_k s_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k}. \qquad \text{(DFP H)}$$

This formula can be extended to higher-rank updates. Let U and V be matrices in $\mathbb{R}^{n\times p}$ for some p between 1 and n. If we define

$$\hat{A} = A + UV^T,$$

then $\hat{A}$ is nonsingular if and only if $(I + V^T A^{-1} U)$ is nonsingular, and in this case we have

$$\hat{A}^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}. \tag{A.28}$$

Figure: Nocedal Wright (A.28)

Applying the same argument directly to the inverse of the Hessian H_k. The updated approximation H_{k+1} must be symmetric and positive definite and must satisfy the secant equation

$$H_{k+1} y_k = s_k.$$

BFGS update:

$$H_{k+1} = (I - \rho_k s_k y_k^{\mathrm{T}}) H_k (I - \rho_k y_k s_k^{\mathrm{T}}) + \rho_k s_k s_k^{\mathrm{T}} \qquad \text{(BFGS)}$$

with $\rho_k = 1/y_k^{\mathrm{T}} s_k$.

How to choose H_0? Depends on situation, information about the problem e.g. starting with an inverse of approximated Hessian calculated by a finite difference at x_0. Otherwise, we can it to identity of diagonal matrix to reflect the scaling of the variables.

BFGS

Given x_0, inverse Hessian approximation H_0, tolerance $\varepsilon > 0$
Set $k = 0$
while $\|\nabla f_k\| > \varepsilon$ **do**
    Compute search direction

$$p_k = -H_k \nabla f_k$$

    $x_{k+1} = x_k + \alpha_k p_k$ where α_k is computed with a line search procedure satisfying Wolfe conditions
    Define $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$
    Compute H_{k+1} using (BFGS)
    $k = k + 1$
end while

- Complexity of each iteration is $\mathcal{O}(n^2)$ plus the cost of function and gradient evaluations.
- There are no $\mathcal{O}(n^3)$ such as linear system solves or matrix-matrix multiplications
- The algorithm is robust and the rate of convergence is superlinear. In many cases it outperforms Netwon method, which while converging quadratically, has higher complexity per iteration (Hessian computation and solve).
- A BFGS version with the Hessian approximation B_k rather than H_k. The update for B_k is obtained by applying Sherman-Morrison-Woodbury formula to (BFGS)

$$B_{k+1} = B_k - \frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k} + \frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k}. \qquad \text{(BFGS B)}$$

An $\mathcal{O}(n^2)$ implementation can be achieved based on updates of LDL^{T} factors of B_k (with possible diagonal modification for stability) but no computational advantage is observed above (BFGS).

- The positive definiteness of H_k is not explicitly forced, but if H_k is positive definite so will be H_{k+1}.
- What happens if at some iteration H_k becomes as poor approximation to the true inverse Hessian. E.g. if $s_k^\mathrm{T} y_k$ is tiny (positive) than the elements of H_{k+1} get very large. It turns out that BFGS has effective self correcting properties, and H_k tends to recover in a few steps. The self correcting properties hold only when adequate line search is performed. In particular Wolfe condiions ensure that the gradients are sampled at points which allow the model m_k to capture the curvature information.
- On the other hand DFP method is less effective in correcting itself.
- DFP and BFGS are dual in the sense that they can be obtained by switching $s \leftrightarrow y, B \leftrightarrow H$.

Implementation

- $\alpha_k = 1$ should always be tried first, because this step length will eventually be accepted (under certain conditions), thereby producing super linear convergence.

- Computational evidence suggest that it is more economical (in terms of function evaluations) to perform fairly inaccurate line search.

- $c_1 = 10^{-4}$, $c_2 = 0.9$ are commonly used with Wolfe conditions.

Heuristic for scaling H_0

Choice $H_0 = \beta I$ is popular, but there is no good strategy for estimating β.

If β is too large, the first step $p_0 = -\beta g_0$ is too long and line search may require many iterations to find a suitable step length α_0.

Heuristic: estimate β after the first step has been computed but before the H_0 update (in step 7) and change the provisional value by setting $H_0 = \frac{s_k^{\mathrm{T}} y_k}{y_k^{\mathrm{T}} y_k} I$. This scaling attempts to approximate scaling with an eigenvalue of the inverse Hessian (from Taylor theorem: the secant equation is satisfied for average Hessian $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau$, $y_k = \bar{G}_k \alpha_k p_k = \bar{G}_k s_k$).

Symmetric rank-1 (SR-1) update

Both BFGS and DFP methods perform a rank-2 update while preserving symmetry and positive definiteness.

Question: Does a rank-1 update exist such that the secant equation is satisfied and the symmetry and definiteness are preserved?

Rank-1 update:

$$B_{k+1} = B_k + \sigma v v^{\mathrm{T}}, \quad \sigma \in \{+1, -1\}$$

and v is chosen such that B_{k+1} satisfies the secant equation

$$y_k = B_{k+1} s_k.$$

Substituting the explicit rank-1 form into the secant equation

$$y_k = B_k s_k + \underbrace{(\sigma v^{\mathrm{T}} s_k)}_{:=\delta} v$$

we see that v must be of the form $v = \delta(y_k - B_k s_k)$.

Substituting $v = \delta(y_k - B_k s_k)$ back into the secant equation we obtain

$$y_k - B_k s_k = \sigma \delta^2 [s_k^{\mathrm{T}}(y_k - B_k s_k)](y_k - B_k s_k)$$

which is satisfied if and only if

$$\sigma = \mathrm{sign}[s_k^{\mathrm{T}}(y_k - B_k s_k)], \quad \delta = \pm|s_k^{\mathrm{T}}(y_k - B_k s_k)|^{-1/2}$$

Hence, the only symmetric rank-1 update satisfying the secant equation is

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^{\mathrm{T}}}{(y_k - B_k s_k)^{\mathrm{T}} s_k}. \qquad \text{(SR-1)}$$

Applying the Sherman-Morrison-Woodbury formula the inverse Hessian update is

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^{\mathrm{T}}}{(s_k - H_k y_k)^{\mathrm{T}} y_k}. \qquad \text{(SR-1)}$$

SR-1 update does not preserve the positive definiteness. It is a drawback for line search methods but could be an asset for trust region as it allows to generate indefinite Hessians.

SR-1 breakdown

The main drawback of SR-1 is that $(s_k - H_k y_k)^{\mathrm{T}} y_k$ can become 0 even for a convex quadratic function i.e. there may be steps where the is no symmetric rank-1 update which satisfies the secant equation.

Three cases:

- $(y_k - B_k s_k)^{\mathrm{T}} s_k \neq 0$, unique symmetric rank-1 update satisfying secant equation exists.
- $y_k = B_k s_k$, then the only update is $B_{k+1} = B_k$.
- $(y_k - B_k s_k)^{\mathrm{T}} s_k = 0$ and $y_k \neq B_k s_k$, there is no symmetric rank-1 update satisfying secant equation.

Remedy: Skipping i.e. apply update only if

$$|(y_k - B_k s_k)^{\mathrm{T}} s_k| \geq r \|s_k\| \|y_k - B_k s_k\|,$$

where $r \in (0, 1)$ is a small number (typically $r = 10^{-8}$), otherwise set $B_{k+1} = B_k$.

SR-1 applicability

- This simple safeguard adequately prevents the breakdown. Recall: for BFGS update skipping is not recommended if the curvature condition $s_k^{\mathrm{T}} y_k > 0$ fails. Because it can occur often by e.g. taking to small step if the line search does not impose the Wolfe conditions. For SR-1 $s_k^{\mathrm{T}}(y_k - B_k s_k) \approx 0$ occurs infrequently as it requires near orthogonality of s_k and $y_k - B_k s_k$ and moreover it implies that $s_k^{\mathrm{T}} \bar{G}_k s_k \approx s_k^{\mathrm{T}} B_k s_k$, where $\bar{G}_k$ is the average Hessian over the last step meaning that the curvature approximation along s_k is essentially already correct.
- The Hessian approximation generated by SR-1 are good, often better than those by BFGS
- When the curvature condition $y_k^{\mathrm{T}} s_k > 0$ cannot be imposed e.g. constraint problems or partially separable functions, indefinite Hessian approximations are desirable as they reflect the indefiniteness of the true Hessian.

Given x_0, B_0, Δ, $\eta \in (0, 10^{-3})$, $r \in (0,1)$ and $\varepsilon > 0$
Set $k = 0$
while $\|\nabla f_k\| > \varepsilon$ **do**
    $s_k = \arg\min_s s^{\mathrm{T}} \nabla f_k + \frac{1}{2} s^{\mathrm{T}} B_k s, \quad$ subject to $\|s\| \le \Delta_k$
    $y_k = \nabla f(x_k + s_k) - \nabla f_k$
    $\rho_k = (f_k - f(x_k + s_k)) / -(s_k^{\mathrm{T}} \nabla f_k + \frac{1}{2} s_k^{\mathrm{T}} B_k s_k)$
    if $\rho_k > \eta$ **then**
        $x_{k+1} = x_k + s_k$
    else
        $x_{k+1} = x_k$ (failed step)
    end if
    Update Δ_k in dependence of ρ_k, $\|s_k\|$ (as in trust-region methods)
    if $|(y_k - B_k s_k)^{\mathrm{T}} s_k| \ge r \|s_k\| \|y_k - B_k s_k\|$ **then**
        Update B_{k+1} using (SR-1) (even if $x_{k+1} = x_k$ to improve bad approximation along s_k)
    else
        $B_{k+1} = B_k$
    end if
    $k = k + 1$
end while

Theorem: Hessian approximation for quadratic function

Let $f : \mathbb{R}^n \to \mathbb{R}$ is the strongly quadratic function $f(x) = b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax$ with A symmetric positive definite. For any starting point x_0 and any symmetric initial matrix H_0, the iterates

$$x_{k+1} = x_k + p_k, \quad p_k = -H_k \nabla f_k,$$

where H_k is updated with (SR-1), converge to the minimiser in at most n steps provided that $(s_k - H_k y_k)^{\mathrm{T}} y_k \neq 0$ for all k. After n steps, if the search directions p_k are linearly independent, $H_n = A^{-1}$.

Proof idea Show inductively that $H_k y_j = s_j$ for all $j = 1, \dots, k-1$ i.e. H_k satisfies the secant equations for all the directions up to now.

For quadratic function the secant equation is satisfied for all previous directions, regardless how the line search is performed. In contrast for BFGS it can only be shown under the assumption that the line search is exact.

Theorem: Hessian approximation for general function

Let $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable with the Hessian bounded and Lipschitz continuous in a neighbourhood of a point $x^\star \in \mathbb{R}^n$ and $\{x_k\}$ a sequence of iterates such that $x_k \to x^\star$. Suppose that

$$|(y_k - B_k s_k)^{\mathrm{T}} s_k| \geq r \|s_k\| \|y_k - B_k s_k\|$$

holds for all k and some $r \in (0, 1)$ and that the steps s_k are uniformly independent (steps do not tend to fall in a subspace of dimension less than n).

Then the matrices B_k generated by the update (SR-1) satisfy

$$\lim_{k \to \infty} \|B_k - \nabla^2 f(x^\star)\| = 0.$$

The Broyden class

Broyden class is a family of updates of the form

$$B_{k+1} = B_k - \frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k^{\mathrm{T}}} + \frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k} + \tau_k (s_k^{\mathrm{T}} B_k s_k) v_k v_k^{\mathrm{T}}, \quad \text{(Broyden)}$$

where τ_k is a scalar parameter and

$$v_k = \frac{y_k}{y_k^{\mathrm{T}} s_k} - \frac{B_k s_k}{s_k^{\mathrm{T}} B_k s_k}.$$

For $\tau_k = 0$ we recover BFGS and for $\tau_k = 1$ we DFP.
Hence we can write (Broyden) as a linear combination of the two

$$B_{k+1} = (1 - \tau_k) B_{k+1}^{\mathrm{BFGS}} + \tau_k B_{k+1}^{\mathrm{DFP}}.$$

Since both BFGS and DFP satisfy secant equation so does the whole Broyden class.
Since BFSG and DFP updates preserve positive definiteness of the Hessian when $s_k^{\mathrm{T}} y_k > 0$, so does the **restricted Broyden class** which is obtained by restricting $0 \le \tau_k \le 1$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ is the strongly convex quadratic function $f(x) = b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax$ with A symmetric positive definite. Let B_0 any symmetric positive matrix and x_0 be any starting point for the iteration

$$x_{k+1} = x_k + p_k, \quad p_k = -B_k^{-1}\nabla f_k,$$

where B_k is updated with (Broyden) with $\tau_k \in [0,1]$.
Denote with $\lambda_1^k \leq \lambda_2^k \leq \cdots \leq \lambda_n^k$ the eigenvalues of

$$A^{1/2}B_k^{-1}A^{1/2}.$$

Then for all k, we have

$$\min\{\lambda_i^k, 1\} \leq \lambda_i^{k+1} \leq \max\{\lambda_i^k, 1\}, \quad i = 1, \ldots, n.$$

The interlacing property does not hold if $\tau_k \notin [0,1]$.

Consequence: The eigenvalues λ_i^k converge monotonically (but not strictly monotonically) to 1, which are the eigenvalues when $B_k = A$. Significantly, the result holds even if the line search is not exact.

So do the best updates belong to the restricted Broyden class?

We recover SR-1 formula for

$$\tau_k = \frac{s_k^{\mathrm{T}} y_k}{s_k^{\mathrm{T}} y_k - s_k^{\mathrm{T}} B_k s_k},$$

which does not belong to the restricted Broyden class as τ_k may fall outside of $[0, 1]$.

It can be shown that for B_0 symmetric positive definite, if for all k $s_k^{\mathrm{T}} y_k > 0$ and $\tau_k > \tau_k^c$, then all B_k generated by (Broyden) remain symmetric and positive definite. Here

$$\tau_k^c = (1 - \mu_k)^{-1} \leq 0, \quad \mu_k = \frac{(y_k^{\mathrm{T}} B_k^{-1} y_k)(s_k^{\mathrm{T}} B_k s_k)}{(y_k^{\mathrm{T}} s_k)^2} \geq 1.$$

When the line search is exact all the methods in the Broyden class with $\tau_k \geq \tau_k^c$ generate the same sequence of iterates, even for nonlinear functions because the directions differ only by length and this is compensated by the exact line search.

Thm: Properties of Broyden class for quadratic function

Let $f : \mathbb{R}^n \to \mathbb{R}$ be the strongly convex quadratic function $f(x) = b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax$ with A symmetric positive definite. Let x_0 be any starting point and B_0 any symmetric positive definite matrix. Assume that α_k is the exact step length and $\tau_k \geq \tau_k^c$ for all k. Then it holds

(i) The iterates are independent of τ_k and converge to the solution in at most n iterations.

(ii) The secant equation is satisfied for all previous search directions

$$B_k s_j = y_j, \quad j = 1, \dots, k-1.$$

(iii) If $B_0 = I$, then the sequence of iterates $\{x_k\}$ is identical to that generated by the conjugate gradient method, in particular the search directions s_k are conjugate

$$s_i^{\mathrm{T}} A s_j = 0, \quad i \neq j.$$

(iv) If n iterations are performed, we have $B_n = A$.

Few comments ...

- The theorem can be slightly generalised to hold if the Hessian approximation remains nonsingular but not necessarily positive definite i.e. τ_k could be smaller than τ_k^c provided the chosen value did not produce singular updated matrix.

- (iii) can be generalised to $B_0 \neq I$, then the Broyden class method is identical to preconditioned conjugate gradient method with the preconditioner B_0.

- The theorem is mainly of theoretical interest as the inexact line search used in practice significantly alters the performance of the methods. This type of analysis however, guided much of the development in quasi-Newton methods.

For general nonlinear objective function, there is no global convergence result for quasi-Newton methods i.e. convergence to a stationary point from any starting point and any suitable Hessian approximation.

Theorem: [BFGS global convergence]
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and x_0 be a starting point for which the level set $\mathcal{L} = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is convex and there exist two positive constants m, M such that

$$m\|z\|^2 \leq z^{\mathrm{T}} \nabla^2 f(x) z \leq M\|z\|^2, \quad \forall z \in \mathbb{R}^n,\ x \in \mathcal{L}.$$

Then for any symmetric positive definite matrix B_0 the sequence $\{x_k\}$ generated by BFGS algorithm (with $\varepsilon = 0$) converges to the miminizer $x^\star$ of f.

This results can be generalised to the restricted Broyden class with $\tau_k \in [0, 1)$ i.e. except for DFP method.

Theorem: Superlinear convergence BFGS

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and the sequence of iterates generated by BFGS algorithm converge to $x^\star \in \mathbb{R}^n$ such that the Hessian $\nabla^2 f$ is Lipschitz continuous at $x^\star$

$$\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \le L\|x - x^\star\|, \quad \forall x \in \mathcal{N}(x_0),\ 0 \le L < \infty,$$

and that it holds

$$\sum_{k=1}^{\infty} \|x_k - x^\star\| < \infty,$$

then x_k converges to $x^\star$ at a superlinear rate.

Theorem: SR-1 trust region convergence

Let $\{x_k\}$ be the sequence of iterates generated by the SR-1 trust region method. Suppose the following conditions hold:

- the sequence $\{x_k\}$ does not terminate, but remains in a closed bounded convex set $\mathcal{D}$ on which f is twice continuously differentiable and in which f has a unique stationary point $x^\star$;
- $\nabla^2 f(x^\star)$ is positive definite and $\nabla^2 f(x)$ is Lipschitz continuous in $\mathcal{N}(x^\star)$;
- the sequence $\{B_k\}$ is bounded in norm;
- $|(y_k - B_k s_k)^T s_k| \geq r\|s_k\|\|y_k - B_k s_k\|, \; r \in (0,1), \; \forall k.$

Then for the sequence $\{x_k\}$ we have $\lim_{k\to\infty} x_k = x^\star$ and

$$\lim_{k\to\infty} \frac{\|x_{k+n+1} - x^\star\|}{\|x_k - x^\star\|} = 0 \quad (n+1\text{-step superlinear rate}).$$

Remark: SR-1 update does not maintain positive definiteness of B_k. In practice B_k can be indefinite at any iteration (trust region bound may continue to be active for arbitraily large k) but it can be shown that (asymptotically) B_k remains positive definite most of the time.