Nonlinear Programming Algorithms Handout

Michael C. Ferris
Computer Sciences Department
University of Wisconsin
Madison, Wisconsin 53706

September 2009

1 Eigenvalues

The eigenvalues of a matrix $A \in \mathbf{C}^{n \times n}$ are the roots of the characteristic polynomial $\phi(\lambda) := \det(\lambda I - A)$. The set of all eigenvalues is called the spectrum and is denoted by $\lambda(A)$. Eigenvectors are vectors x satisfying $Ax = \lambda x$ for some $\lambda \in \lambda(A)$.

If $A \in \mathbf{R}^{n \times n}$ is symmetric then

$$A = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

where $Q \in \mathbf{R}^{n \times n}$ is orthogonal (basis of eigenvectors) and $\Lambda \in \mathbf{R}^{n \times n}$ is diagonal (each entry called an eigenvalue). Note that in this case the eigenvalues and eigenvectors may be taken as real. The above factorization is called a spectral decomposition.

For real symmetric matrices, the eigenvalues are assumed to be ordered using $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

2 Singular Values

Orthogonal matrices satisfy $U^T U = U U^T = I$. If $A \in \mathbf{R}^{m \times n}$ then there exist orthogonal matrices $U = [u_1, \ldots, u_m] \in \mathbf{R}^{m \times m}$ and $V = [v_1, \ldots, v_n] \in \mathbf{R}^{n \times n}$ such that

$$U^T A V = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbf{R}^{m \times n}$$

where $p = \min\{m, n\}$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$. This is often written as $A = USV^T$ and is termed the singular value decomposition. Singular values are non-negative; $\sigma_i(A)$ denote the ith largest singular value of A.

If we define r by

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0$$

then $\operatorname{rank}(A) = r$, $\ker(A) = span\{v_{r+1}, \ldots, v_n\}$, $\operatorname{im}(A) = span\{u_1, \ldots, u_r\}$ and

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

How are eigenvalues and singular values related?

$$\sigma(A) = +\sqrt{\lambda(AA^T)}$$

Definition 1 *A matrix A is positive definite if $x^T Ax > 0$ for all $x \neq 0$. A matrix A is positive semidefinite if $x^T Ax \geq 0$ for all x.*

When A is a symmetric positive definite matrix, then singular values and eigenvalues coincide (take $U = V = Q$ and $S = \Lambda$), and thus

$$\sigma_1(A) = \text{largest eigenvalue of } A$$

$$\sigma_n(A) = \text{smallest eigenvalue of } A$$

If A is square, then

$$\sigma_n(A) \leq \min_i |\lambda_i| \leq \max_i |\lambda_i| \leq \sigma_1(A)$$

Lemma 2 *For a symmetric matrix $A \in \mathbf{R}^{n \times n}$, the following are equivalent*

1. *A is positive definite*
2. *The (n) leading principal subdeterminants of A are strictly positive*
3. *All principal subdeterminants of A are strictly positive*
4. *The eigenvalues of A are strictly positive*

The matrix

$$\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$$

shows symmetry to be necessary for this result.

3 Matrix Norms

Suppose $A \in \mathbf{R}^{m \times n}$. $\|A\|_{\alpha,\beta} := \sup_{x \neq 0} \|Ax\|_\beta \|x\|_\alpha$ In particular, when $\alpha = \beta = p$, this is called the p-norm of A. The Frobenius norm is defined by $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2}$. Various properties of matrix norms detail in GVL (pp 54).

Interesting facts:

$$\|A\|_2 = \sigma_1$$

$$\|A\|_F^2 = \sum_{i=1}^{p} \sigma_i^2, p = \min(m,n)$$

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2$$

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |A_{ij}|$$

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |A_{ij}|$$

$$\max_{i,j} |A_{ij}| \leq \|A\|_2 \leq \sqrt{mn}\max_{i,j} |A_{ij}|$$

$$\|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty}$$

$$\min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_n$$

Lemma 3 *Let A be a real symmetric matrix. For each $x \in \mathbf{R}^n$*

$$\lambda_n \|x\|^2 \leq x^T A x \leq \lambda_1 \|x\|^2$$

If A is also positive definite, then $\|A\| = \lambda_1$.

Proof Since $A = Q\Lambda Q^T$ (Q invertible), we let $x = Qy$ for some y. Then

$$x^T A x = y^T Q^T A Q y = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2.$$

Also

$$\|y\|^2 = y^T y = x^T Q Q^T x = \|x\|^2.$$

Thus,

$$\lambda_n \|x\|^2 = \lambda_n \sum_{i=1}^{n} y_i^2 \leq \sum_{i=1}^{n} \lambda_i y_i^2 (= x^T A x) \leq \lambda_1 \sum_{i=1}^{n} y_i^2 = \lambda_1 \|x\|^2$$

Finally, $Ax = Q\Lambda Q^T Q y = Q\Lambda y$, so

$$\|Ax\|^2 = y^T \Lambda Q^T Q \Lambda y = \sum_{i=1}^{n} \lambda_i^2 y_i^2 \leq \lambda_1^2 \|y\|^2 = \lambda_1^2 \|x\|^2,$$

the inequality following from the positive definite assumption. Hence $\frac{\|Ax\|^2}{\|x\|^2} \leq \lambda_1^2$ so $\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \leq \lambda_1$. But the supremum is attained at an eigenvector of A corresponding to λ_1. □

Definition 4 *Suppose $A \in \mathbf{R}^{n\times n}$ is nonsingular. The* condition number *of A, $\kappa(A)$, is defined to be $\|A\|\,\|A^{-1}\|$.*

The condition number of a matrix depends on the underlying norm. If A is singular, then $\kappa(A) = \infty$. Note that

$$\kappa_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2 = \frac{\sigma_1(A)}{\sigma_n A}.$$

Corollary 5 *If A is real symmetric positive definite, then*

$$\lambda_1^{-1} \|x\|^2 \leq x^T A^{-1} x \leq \lambda_n^{-1} \|x\|^2$$

$\|A^{-1}\| = \lambda_n^{-1}$ *and* $\kappa(A) = \frac{\lambda_1}{\lambda_n}$.

Proof The eigenvalues of A^{-1} are simply λ_i^{-1}. □

Definition 6 *Suppose $A \in \mathbf{C}^{n\times n}$. The* spectral radius *of A, $\rho(A)$, is defined as the maximum of $|\lambda_1|, \ldots, |\lambda_n|$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.*

Note that if $A \in \mathbf{C}^{n\times n}$ and λ is any eigenvalue of A with eigenvector $u \neq 0$, then $\|Au\| = |\lambda| \|u\|$, so that $|\lambda| \leq \|A\|$. Hence $\rho(A) \leq \|A\|$.

Lemma 7 *Let $A \in \mathbf{C}^{n\times n}$. Then $\lim_{k\to\infty} A^k = 0$ if and only if $\rho(A) < 1$.*

Lemma 8 (Neumann) *Let $E \in \mathbf{R}^{n\times n}$ and $\rho(E) < 1$. Then $I - E$ is nonsingular and*

$$I - E^{-1} = \lim_{k\to\infty} \sum_{i=0}^{k} E^i$$

Proof Since $\rho(E) < 1$, the maximum eigenvalue of E must be less than 1 in modulus and hence $I - E$ is invertible. Furthermore

$$(I - E)(I + \cdots + E^{k-1}) = I - E^k$$

so that

$$I + E + \cdots + E^{k-1} = (I - E)^{-1} - (I - E)^{-1} E^k$$

The result now follows in the limit as $k \to \infty$ from the above lemma. □

4 Results from Matrix Theory

Lemma 9 (Debreu) *Suppose $H \in \mathbf{R}^{n\times n}$ and $A \in \mathbf{R}^{m\times n}$. The following are equivalent:*

(a) *$Az = 0$ and $z \neq 0$ implies $\langle z, Hz\rangle > 0$*

(b) *There exists $\bar{\gamma}$ such that $H + \bar{\gamma} A^T A$ is positive definite*

Remark If (b) holds and $\gamma \geq \bar{\gamma}$ then for any z,

$$\begin{aligned} \left\langle z, (H + \gamma A^T A) z \right\rangle &= \left\langle z, (H + \bar{\gamma} A^T A) z \right\rangle + (\gamma - \bar{\gamma}) \|Az\|^2 \\ &\geq \left\langle z, (H + \bar{\gamma} A^T A) z \right\rangle \end{aligned}$$

so that $H + \gamma A^T A$ is also positive definite.

Proof

(b) $\Rightarrow$ (a) If $Az = 0$ and $z \neq 0$, then $0 < \langle z, Hz \rangle = \left\langle z, (H + \bar{\gamma} A^T A) z \right\rangle$.

(a) $\Rightarrow$ (b) If (b) were false, then there would exist z^k of norm 1, with

$$\left\langle z^k, (H + k A^T A) z^k \right\rangle \leq 0$$

Without loss of generality, we can assume that $z^k \to z^0$. Now for each k

$$0 \geq \left\langle z^k, (k^{-1} H + A^T A) z^k \right\rangle \to \left\| A z^0 \right\|^2 \geq 0$$

so $Az^0 = 0$. But

$$0 \geq \left\langle z^k, H z^k \right\rangle + k \left\| A z^k \right\|^2 \geq \left\langle z^k, H z^k \right\rangle \to \left\langle z^0, H z^0 \right\rangle$$

and this contradicts (a), since $\|z^0\| = 1$. $\square$

Lemma 10 (Schur) *Suppose that*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{1}$$

with D nonsingular. Then M is nonsingular if and only if $S = A - BD^{-1}C$ is nonsingular, and in that case

$$M^{-1} = \begin{pmatrix} S^{-1} & -S^{-1} B D^{-1} \\ -D^{-1} C S^{-1} & D^{-1} + D^{-1} C S^{-1} B D^{-1} \end{pmatrix} \tag{2}$$

Remark S is the Schur complement of D in M. We can also show that $\det M = \det S \det D$ since

$$\begin{aligned} \det M &= \det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det \left(\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I & 0 \\ -C & D^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ DC & D \end{pmatrix} \right) \\ &= \det \begin{pmatrix} A - BD^{-1}C & BD^{-1} \\ 0 & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ DC & D \end{pmatrix} \\ &= \det A - BD^{-1}C \det D \end{aligned}$$

Proof If S is nonsingular, then just multiply the expressions given in (1) and (2) to obtain I. Thus assume M is nonsingular. Write its inverse as

$$M^{-1} = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

Since $MM^{-1} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$ we have

$$\begin{aligned} I &= AE + BG & (3) \\ 0 &= AF + BH & (4) \\ 0 &= CE + DG & (5) \\ I &= CF + DH & (6) \end{aligned}$$

From (5) we have $G = -D^{-1}CE$, and substituting this in (3) yields $(A - BD^{-1}C)E = I$. This implies that $S = A - BD^{-1}C$ is nonsingular and that $E = S^{-1}$. Using (5) again we have that $G = -D^{-1}CS^{-1}$. From (6) we have $H = D^{-1}(I - CF)$, and putting this in (4) yields

$$AF + BD^{-1} - BD^{-1}CF = 0$$

that is

$$SF + BD^{-1} = 0$$

which implies $F = -S^{-1}BD^{-1}$. □

5 Matrix Norms

Suppose $A \in \mathbf{R}^{m\times n}$. $\|A\|_{\alpha,\beta} := \sup_{x\neq 0} \|Ax\|_\beta \|x\|_\alpha$ In particular, when $\alpha = \beta = p$, this is called the p-norm of A. The Frobenius norm is defined by $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2}$. Various properties of matrix norms detaile in GVL (pp 54).

Interesting facts: The two norm of A is the square root of the largest eigenvalue of A^TA.

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2$$

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |A_{ij}|$$

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |A_{ij}|$$

$$\max_{i,j} |A_{ij}| \leq \|A\|_2 \leq \sqrt{mn}\max_{i,j} |A_{ij}|$$

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$$

The condition number of a matrix depends on the underlying norm, but is defined for a square matrix by

$$\kappa(A) := \|A\| \left\|A^{-1}\right\|$$

If A is singular, then $\kappa(A) = \infty$. Note that

$$\kappa_2(A) = \|A\|_2 \|invA\|_2 = \frac{\sigma_1(A)}{\sigma_n A}$$

where $\sigma_i(A)$ is the ith largest singular value of A. Singular values are non-negative. How are they related to eigenvalues?

$$\|A\|_F^2 = \sum_{i=1}^{p} \sigma_i^2, p = \min(m, n)$$

$$\|A\|_2 = \sigma_1$$

$$\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_n$$

6 Positive (Semi)-Definite Matrices

Definition 11 *A matrix A is positive definite if $x^T Ax > 0$ for all $x \neq 0$. A matrix A is positive semidefinite if $x^T Ax \geq 0$ for all x.*

Theorem 12 *See Bertsekas*

7 Strong Convexity

Let $f: \Omega \to \mathbf{R}$, $h: \Omega \to \mathbf{R}$ where Ω is an open convex set.

Definition 13 *f is* strongly convex *(ρ) on Ω if $\exists \rho > 0$ such that $\forall x, y \in \Omega, \lambda \in [0, 1]$*

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y) - \frac{\rho}{2}\lambda(1 - \lambda) \|x - y\|^2$$

Definition 14 *h is* strongly monotone *(ρ) on Ω if $\exists \rho > 0$ such that $\forall x, y \in \Omega$*

$$\langle h(x) - h(y), x - y\rangle \geq \rho \|x - y\|^2$$

Theorem 15 (Strong convexity) *If f is continuously differentiable on Ω then the following are equivalent:*

(a) *f is strongly convex (ρ) on Ω*

(b) *For all x, $y \in \Omega$, $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + (\rho/2) \|x - y\|^2$*

(c) *∇f is strongly monotone (ρ) on Ω*

If f is twice continuously differentiable on Ω, then

(d) *For all x,y,z $\in \Omega$, $\langle x - y, \nabla^2 f(z)(x - y)\rangle \geq \rho \|x - y\|^2$*

is equivalent to the above.

Proof We show (a) $\iff$ (b) $\iff$ (c).

(a) $\Rightarrow$ (b) The hypothesis gives

$$\frac{f(x+\lambda(y-x))-f(x)}{\lambda} \leq f(y)-f(x)-\frac{\rho}{2}(1-\lambda)\|x-y\|^2$$

so taking the limit as $\lambda \to 0$

$$\langle \nabla f(x), y-x\rangle \leq f(y)-f(x)-\frac{\rho}{2}\|x-y\|^2$$

(b) $\Rightarrow$ (c) Applying (b) twice gives

$$\begin{aligned} f(y) &\geq f(x)+\langle \nabla f(x), y-x\rangle+\frac{\rho}{2}\|x-y\|^2 \\ f(x) &\geq f(y)+\langle \nabla f(y), x-y\rangle+\frac{\rho}{2}\|x-y\|^2 \end{aligned}$$

Adding these inequalities gives

$$f(y)+f(x) \geq f(x)+f(y)+\langle \nabla f(x)-\nabla f(y), y-x\rangle+\rho\|x-y\|^2$$

from where the result follows.

(c) $\Rightarrow$ (b) The hypothesis gives

$$\int_0^1 \langle \nabla f(x+t(y-x))-\nabla f(x), y-x\rangle\, dt \geq \int_0^1 \rho t\|x-y\|^2\, dt$$

which implies the result.

(b) $\Rightarrow$ (a) Letting $y=u$ and $x=(1-\lambda)u+\lambda v$ in (b) gives

$$f(u) \geq f((1-\lambda)u+\lambda v)+\langle \nabla f((1-\lambda)u+\lambda v), \lambda(u-v)\rangle+\frac{\rho}{2}\|\lambda(u-v)\|^2 \tag{7}$$

Also letting $y=v$ and $x=(1-\lambda)u+\lambda v$ in (b) implies

$$f(v) \geq f((1-\lambda)u+\lambda v)+\langle \nabla f((1-\lambda)u+\lambda v), (1-\lambda)(v-u)\rangle+\frac{\rho}{2}\|(1-\lambda)(v-u)\|^2 \tag{8}$$

Adding $(1-\lambda)$ times (7) to λ times (8) gives the required result.
To complete the proof, we assume that f is twice continuously differentiable on Ω.

(d) $\Rightarrow$ (c) This follows from the hypothesis since

$$\begin{aligned} \langle \nabla f(x)-\nabla f(y), x-y\rangle &= \left\langle \int_0^1 \nabla^2 f(y+t(x-y))(x-y)dt, x-y\right\rangle \\ &\geq \rho\|x-y\|^2 \end{aligned}$$

(c) $\Rightarrow$ (d) Let x, y, $z \in \Omega$. Then $z+\lambda(x-y) \in \Omega$ for sufficiently small λ, so

$$\begin{aligned} \left\langle x-y, \nabla^2 f(z+\lambda(x-y))(x-y)\right\rangle &= \frac{\langle x-y, \nabla f(z+\lambda(x-y))-\nabla f(z)\rangle}{\lambda}+o(1) \\ &\geq \rho\|x-y\|^2+o(1) \end{aligned}$$

The result follows in the limit as $\lambda \to 0$. □

8 Lipschitz Continuity

Definition 16 *f is* Lipschitz continuous *(ρ) on Ω if $\exists \rho > 0$ such that $\forall x, y \in \Omega$*

$$\|f(y) - f(x)\| \leq \rho \|y - x\|$$

Lemma 17 (Lipschitz continuity) *Consider the following statements:*

(a) *∇f is Lipschitz continuous on Ω with constant ρ*

(b) *For all x, $y \in \Omega$, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + (\rho/2) \|x - y\|^2$*

(c) *For all x, $y \in \Omega$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \rho \|x - y\|^2$*

Then (a) implies (b) implies (c). If f is twice continuously differentiable, then (c) implies

(d) *For all x, $y \in \Omega$, $\langle y - x, \nabla^2 f(z)(y - x) \rangle \leq \rho \|y - x\|^2$*

If $\langle x - y, \nabla^2 f(x)(y - x) \rangle \geq \gamma \|y - x\|^2$ for some γ (not necessarily positive), then (d) implies (a), possibly with a different constant ρ.

Proof

(a) $\Rightarrow$ (b)

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \\ &\leq \|y - x\| \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, dt \\ &\leq \rho \|y - x\|^2 \int_0^1 t dt \\ &= \frac{\rho}{2} \|y - x\|^2 \end{aligned}$$

(b) $\Rightarrow$ (c) Invoking (b) twice gives

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\rho}{2} \|x - y\|^2 \\ f(x) &\leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\rho}{2} \|x - y\|^2 \end{aligned}$$

Adding these inequalities gives

$$f(y) + f(x) \leq f(x) + f(y) + \langle \nabla f(x) - \nabla f(y), y - x \rangle + \rho \|x - y\|^2$$

from where (c) follows.

(c) $\Rightarrow$ (d) It follows from (c) that

$$\langle \nabla f(x + \lambda(y - x)) - \nabla f(x), \lambda(y - x) \rangle \leq \rho \lambda^2 \|y - x\|^2$$

If we divide both sides by λ^2, (d) then follows in the limit as $\lambda \to 0$.

(d) $\Rightarrow$ (a) Let $\delta := 2\max\{-\gamma, 0\} + \rho$. We first show that $\|\nabla^2 f(z)\| \leq \delta$. Note that

$$\left\|\nabla^2 f(z)\right\| = \sup_{\|y\|=1} \left\|\nabla^2 f(z)y\right\| = \sup_{\|x\|=1, \|y\|=1} \left\|\left\langle x, \nabla^2 f(z)y\right\rangle\right\|$$

However,

$$\begin{aligned}
\left\langle x, \nabla^2 f(z)y\right\rangle &= \frac{1}{2}\left\langle x-y, \nabla^2 f(z)(y-x)\right\rangle + \left\langle x, \nabla^2 f(z)x\right\rangle + \left\langle y, \nabla^2 f(z)y\right\rangle \\
&\leq \frac{1}{2}\{-\gamma\|x-y\|^2 + \rho\|x\|^2 + \|y\|^2\} \\
&\leq \frac{1}{2}\{\max\{-\gamma, 0\}(\|x\|^2 + 2\|x\|\|y\| + \|y\|^2) + \rho(\|x\|^2 + \|y\|^2)\}
\end{aligned}$$

Hence, $\|\nabla^2 f(z)\| \leq 2\max\{-\gamma, 0\} + \rho$, as required. The Lipschitz continuity now follows easily since

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| &= \left\|\int_0^1 \nabla^2 f(x + t(y-x))(y-x)dt\right\| \\
&\leq \delta\|y-x\|
\end{aligned}$$

□