Elementary Gradient-Based Parameter Estimation

JULIUS O. SMITH III

Center for Computer Research in Music and Acoustics (CCRMA)
Department of Music, Stanford University, Stanford, California 94305 USA

Abstract

This section defines some of the basic terms involved in optimization techniques known as *gradient descent* and *Newton's method*. Terms defined include *metric space*, *linear space*, *norm*, *pseudo-norm*, *normed linear space*, *Banach space*, *Lp space*, *Hilbert space*, *functional*, *convex norm*, *concave norm*, *local minimizer*, *global minimizer*, and *Taylor series expansion*.

Contents

1 Vector Space Concepts

Definition. A set X of objects is called a *metric space* if with any two points p and q of X there is associated a real number $d(p,q)$, called the distance from p to q, such that (a) $d(p,q) > 0$ if $p \neq q$; $d(p,p) = 0$, (b) $d(p,q) = d(q,p)$, (c) $d(p,q) \leq d(p,r) + d(r,q)$, for any $r \in X$ [6].

Definition. A *linear space* is a set of "vectors" X together with a field of "scalars" $\mathcal{S}$ with an addition operation $+ : X \times X \mapsto X$, and a multiplication opration $\cdot$ taking $\mathcal{S} \times X \mapsto X$, with the following properties: If x, y, and z are in X, and α, β are in $\mathcal{S}$, then

1. $x + y = y + x$.
2. $x + (y + z) = (x + y) + z$.
3. There exists $\emptyset$ in X such that $0 \cdot x = \emptyset$ for all x in X.
4. $\alpha(\beta x) = (\alpha\beta)x$.
5. $(\alpha + \beta)x = \alpha x + \beta x$.
6. $1 \cdot x = x$.
7. $\alpha(x + y) = \alpha x + \alpha y$.

The element $\emptyset$ is written as 0 thus coinciding with the notation for the real number zero. A linear space is sometimes be called a linear vector space, or a vector space.

Definition. A *normed linear space* is a linear space X on which there is defined a real-valued function of $x \in X$ called a *norm*, denoted $\| x \|$, satisfying the following three properties:

1. $\| x \| \geq 0$, and $\| x \| = 0 \Leftrightarrow x = 0$.
2. $\| cx \| = |c| \cdot \| x \|$, c a scalar.
3. $\| x_1 + x_2 \| \leq \| x_1 \| + \| x_2 \|$.

The functional $\| x - y \|$ serves as a distance function on X, so a normed linear space is also a metric space.

Note that when X is the space of continuous complex functions on the unit circle in the complex plane, the norm of a function is not changed when multiplied by a function of modulus 1 on the unit circle. In signal processing terms, the norm is insensitive to multiplication by a unity-gain allpass filter (also known as a Blaschke product).

Definition. A *pseudo-norm* is a real-valued function of $x \in X$ satisfying the following three properties:

1. $\| x \| \geq 0$, and $x = 0 \implies \| x \| = 0$.
2. $\| cx \| = |c| \cdot \| x \|$, c a scalar.
3. $\| x_1 + x_2 \| \leq \| x_1 \| + \| x_2 \|$.

A pseudo-norm differs from a norm in that the pseudo-norm can be zero for nonzero vectors (functions).

Definition. A *Banach Space* is a *complete* normed linear space, that is, a normed linear space in which every Cauchy sequence[1] converges to an element of the space.

Definition. A function $H(e^{j\omega})$ is said to belong to the space L^p if

$$\int_{-\pi}^{\pi} \left|H(e^{j\omega})\right|^p \frac{d\omega}{2\pi} < \infty.$$

Definition. A function $H(e^{j\omega})$ is said to belong to the space H^p if it is in L^p and if its analytic continuation $H(z)$ is analytic for $|z| < 1$. $H(z)$ is said to be in H^{-p} if $H(z^{-1}) \in H^p$.

Theorem. (Riesz-Fischer) The L^p spaces are complete. **Proof.** See Royden [5], p. 117.

Definition. A Hilbert space is a Banach space with a symmetric bilinear inner product $< x, y >$ defined such that the inner product of a vector with itself is the square of its norm $< x, x >= \| x \|^2$.

1.1 Specific Norms

The L^p *norms* are defined on the space L^p by

$$\| F \|_p \triangleq \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left|F(e^{j\omega})\right|^p \frac{d\omega}{2\pi} \right)^{1/p}, \quad p \geq 1. \tag{1}$$

L^p norms are technically pseudo-norms; if functions in L^p are replaced by equivalence classes containing all functions equal almost everywhere, then a norm is obtained.

Since all practical desired frequency responses arising in digital filter design problems are bounded on the unit circle, it follows that $\{H(e^{j\omega})\}$ forms a Banach space under any L^p norm.

The *weighted* L^p *norms* are defined by

$$\| F \|_p \triangleq \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left|F(e^{j\omega})\right|^p W(e^{j\omega}) \frac{d\omega}{2\pi} \right)^{\frac{1}{p}}, \quad p \geq 1, \tag{2}$$

where $W(e^{j\omega})$ is real, positive, and integrable. Typically, $\int W = 1$. If $W(e^{j\omega}) = 0$ for a set of nonzero measure, then a pseudo-norm results.

The case $p = 2$ gives the popular *root mean square norm*, and $\| \cdot \|_2^2$ can be interpreted as the total energy of F in many physical contexts.

An advantage of working in L^2 is that the norm is provided by an *inner product*,

$$\langle H, G \rangle \triangleq \int_{-\pi}^{\pi} H(e^{j\omega})\overline{G(e^{j\omega})} \frac{d\omega}{2\pi}.$$

The norm of a vector $H \in L^2$ is then given by

$$\| H \| \triangleq \sqrt{\langle H, H \rangle}.$$

[1] A sequence $H_n(e^{j\omega})$ is said to be a *Cauchy sequence* if for each $\epsilon > 0$ there is an N such that $\| H_n(e^{j\omega}) - H_m(e^{j\omega}) \| < \epsilon$ for all n and m larger than N.

As p approaches infinity in Eq. (1), the error measure is dominated by the largest values of $|F(e^{j\omega})|$. Accordingly, it is customary to define

$$\| F \|_\infty \stackrel{\Delta}{=} \max_{-\pi<\omega\leq\pi} \left|F(e^{j\omega})\right|, \tag{3}$$

and this is often called the *Chebyshev* or *uniform norm*.

Suppose the L^1 norm of $F(e^{j\omega})$ is finite, and let

$$f(n) \stackrel{\Delta}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} F(e^{j\omega})e^{j\omega n}\frac{d\omega}{2\pi}$$

denote the Fourier coefficients of $F(e^{j\omega})$. When $F(e^{j\omega})$ is a filter frequency response, $f(n)$ is the corresponding *impulse response*. The filter F is said to be *causal* if $f(n) = 0$ for $n < 0$.

The norms for impulse response sequences $\| f \|_p$ are defined in a manner exactly analogous with the frequency response norms $\| F \|_p$, viz.,

$$\| f \|_p \stackrel{\Delta}{=} \left(\sum_{n=-\infty}^{\infty} |f(n)|^p\right)^{\frac{1}{p}}.$$

These time-domain norms are called *l^p norms*.

The L^p and l^p norms are *strictly concave* functionals for $1 < p < \infty$ (see below).

By Parseval's theorem, we have $\| F \|_2 = \| f \|_2$, *i.e.*, the L^p and l^p norms are the same for $p = 2$.

The *Frobenious norm* of an $m \times n$ matrix A is defined as

$$\| A \|_F \stackrel{\Delta}{=} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

That is, the Frobenious norm is the L^2 norm applied to the elements of the matrix. For this norm there exists the following.

Theorem. The unique $m \times n$ rank k matrix B which minimizes $\| A - B \|_F$ is given by $U\Sigma_k V^*$, where $A = U\Sigma V^*$ is a singular value decomposition of A, and Σ_k is formed from Σ by setting to zero all but the k largest singular values.

Proof. See Golub and Kahan [3].

The *induced norm* of a matrix A is defined in terms of the norm defined for the vectors $\underline{x}$ on which it operates,

$$\| A \| \stackrel{\Delta}{=} \sup_{\underline{x}} \frac{\| A\underline{x} \|}{\| \underline{x} \|}$$

For the L^2 norm, we have

$$\| A \|_2^2 = \sup_{\underline{x}} \frac{\underline{x}^T A^T A \underline{x}}{\underline{x}^T \underline{x}},$$

and this is called the *spectral norm* of the matrix A.

The *Hankel matrix* corresponding to a time series f is defined by $\Gamma(f)[i,j] \stackrel{\Delta}{=} f(i+j)$, *i.e.*,

$$\Gamma(f) \stackrel{\Delta}{=} \left(\begin{array}{cccc} f(0) & f(1) & f(2) & \cdots \\ f(1) & f(2) & & \\ f(2) & & & \\ \vdots & & & \end{array} \right).$$

Note that the Hankel matrix involves only causal components of the time series.

The *Hankel norm* of a filter frequency response is defined as the spectral norm of the Hankel matrix of its impulse response,

$$\left\| F(e^{j\omega}) \right\|_H \stackrel{\Delta}{=} \| \Gamma(f) \|_2 .$$

The Hankel norm is truly a norm only if $H(z) \in H^{-p}$, *i.e.*, if it is causal. For noncausal filters, it is a pseudo-norm.

If F is strictly stable, then $|F(e^{j\omega})|$ is finite for all ω, and all norms defined thus far are finite. Also, the Hankel matrix $\Gamma(f)$ is a bounded linear operator in this case.

The Hankel norm is bounded below by the L^2 norm, and bounded above by the L^∞ norm [1],

$$\| F \|_2 \leq \| F \|_H \leq \| F \|_\infty ,$$

with equality iff F is an allpass filter (*i.e.*, $|F(e^{j\omega})|$ constant).

2 Concavity (Convexity)

Definition. A set S is said to be *concave* if for every vector x and y in S, $\lambda x + (1-\lambda)y$ is in S for all $0 \leq \lambda \leq 1$. In other words, all points on the line between two points of S lie in S.

Definition. A *functional* is a mapping from a vector space to the real numbers $\Re$.
Thus, for example, every *norm* is a functional.

Definition. A *linear functional* is a functional f such that for each x and y in the linear space X, and for all scalars α and β, we have $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$.

Definition. The *norm of a linear functional* f is defined on the normed linear space X by

$$\| f \| \stackrel{\Delta}{=} \sup_{x \in X} \frac{|f(x)|}{\| x \|}.$$

Definition. A functional f defined on a concave subset S of a vector space X is said to be *concave* on S if for every vector x and y in S,

$$\lambda f(x) + (1-\lambda) f(y) \geq f\left(\lambda x + (1-\lambda)y\right), \qquad 0 \leq \lambda \leq 1.$$

A concave functional has the property that its values along a line segment lie below or on the line between its values at the end points. The functional is *strictly concave* on S if strict inequality holds above for $\lambda \in (0,1)$. Finally, f is *uniformly concave* on S if there exists $c > 0$ such that for all $x, y \in S$,

$$\lambda f(x) + (1-\lambda) f(y) - f\left(\lambda x + (1-\lambda)y\right) \geq c\lambda(1-\lambda) \| x - y \|^2 , \qquad 0 \leq \lambda \leq 1.$$

We have

$$\text{Uniformly Concave} \implies \text{Strictly Concave} \implies \text{Concave}$$

Definition. A *local minimizer* of a real-valued function $f(x)$ is any x^* such that $f(x^*) < f(x)$ in some neighborhood of x.

Definition. A *global minimizer* of a real-valued function $f(x)$ on a set S is any $x^* \in S$ such that $f(x^*) < f(x)$ for all $x \in S$.

Definition. A *cluster point* x of a sequence x_n is any point such that every neighborhood of x contains at least one x_n.

Definition. The *concave hull* of a set S in a metric space is the smallest concave set containing S.

2.1 Concave Norms

A desirable property of the error norm minimized by a filter-design technique is concavity of the error norm with respect to the filter coefficients. When this holds, the error surface "looks like a bowl," and the *global minimum* can be found by iteratively moving the parameters in the "downhill" (negative gradient) direction. The advantages of concavity are evident from the following classical results.

Theorem. If X is a vector space, S a concave subset of X, and f a concave functional on S, then any local minimizer of f is a global minimizer of f in S.

Theorem. If X is a normed linear space, S a concave subset of X, and f a *strictly* concave functional on S, then f has *at most* one minimizer in S.

Theorem. Let S be a closed and bounded subset of $\Re^n$. If $f : \Re^n \mapsto \Re^1$ is *continuous* on S, then f has *at least* one minimizer in S.

Theorem (2.1) bears directly on the existence of a solution to the general filter design problem in the frequency domain. Replacing "closed and bounded" with "compact", it becomes true for a functional on an arbitrary metric space (Rudin [6], Thm. 14). (In $\Re^n$, "compact" is equivalent to "closed and bounded" [5].) Theorem (2.1) implies only compactness of $\hat{\Theta} = \{\hat{b}_0, \dots, \hat{b}_{n_b}, \hat{a}_1, \dots, \hat{a}_{n_a}\}$ and continuity of the error norm $J(\hat{\theta})$ on $\hat{\Theta}$ need to be shown to prove existence of a solution to the general frequency-domain filter design problem.

3 Gradient Descent

Concavity is valuable in connection with the *Gradient Method* of minimizing $J(\hat{\theta})$ with respect to $\hat{\theta}$.

Definition. The *gradient* of the error measure $J(\hat{\theta})$ is defined as the $\hat{N} \times 1$ column vector

$$J'(\hat{\theta}) \triangleq \frac{\partial J(\theta)}{\partial \theta}(\hat{\theta}) \triangleq \left[\frac{\partial}{\partial \theta} J(\theta) b_0 \left(\hat{b}_0\right), \dots, \frac{\partial}{\partial \theta} J(\theta) b_{n_b} \left(\hat{b}_{n_b}\right), \frac{\partial}{\partial \theta} J(\theta) a_1 \left(\hat{a}_1\right), \dots, \frac{\partial}{\partial \theta} J(\theta) a_{n_a} \left(\hat{a}_{n_a}\right)\right]^T.$$

Definition. The *Gradient Method* (Cauchy) is defined as follows.

Given $\hat{\theta}_0 \in \hat{\Theta}$, compute

$$\hat{\theta}_{n+1} = \hat{\theta}_n - t_n J'(\hat{\theta}_n), \qquad n = 1, 2, \dots ,$$

where $J'(\hat{\theta}_n)$ is the *gradient* of J at $\hat{\theta}_n$, and $t_n \in \Re$ is chosen as the smallest nonnegative local minimizer of

$$\Phi_n(t) \triangleq J\left(\hat{\theta}_n - tJ'(\hat{\theta}_n)\right).$$

Cauchy originally proposed to find the value of $t_n \geq 0$ which gave a global minimum of $\Phi_n(t)$. This, however, is not always feasible in practice.

Some general results regarding the Gradient Method are given below.

Theorem. If $\hat{\theta}_0$ is a local minimizer of $J(\hat{\theta})$, and $J'(\hat{\theta}_0)$ exists, then $J'(\hat{\theta}_0) = 0$.

Theorem. The gradient method is a *descent* method, *i.e.*, $J(\hat{\theta}_{n+1}) \leq J(\hat{\theta}_n)$.

Definition. $J : \hat{\Theta} \to \Re^1$, $\hat{\Theta} \subset \Re^{\hat{N}}$, is said to be in the class $\mathcal{C}_k(\hat{\Theta})$ if all kth order partial derivatives of $J(\hat{\theta})$ with respect to the components of $\hat{\theta}$ are continuous on $\hat{\Theta}$.

Definition. The *Hessian* $J''(\hat{\theta})$ of J at $\hat{\theta}$ is defined as the matrix of second-order partial derivatives,

$$J''(\hat{\theta})[i,j] \triangleq \frac{\partial^2 J(\theta)}{\partial\theta[i]\partial\theta[j]}(\hat{\theta}),$$

where $\theta[i]$ denotes the ith component of θ, $i = 1, \dots , \hat{N} = n_a + n_b + 1$, and $[i, j]$ denotes the matrix entry at the ith row and jth column.

The Hessian of every element of $\mathcal{C}_2(\hat{\Theta})$ is a *symmetric matrix* [7]. This is because continuous second-order partials satisfy

$$\frac{\partial^2}{\partial x_1 \partial x_2} = \frac{\partial^2}{\partial x_2 \partial x_1}.$$

Theorem. If $J \in \mathcal{C}_1(\hat{\Theta})$, then any cluster point $\hat{\theta}_\infty$ of the gradient sequence $\hat{\theta}_n$ is necessarily a *stationary point*, i.e., $J'(\hat{\theta}_\infty) = 0$.

Theorem. Let $\bar{\hat{\Theta}}$ denote the concave hull of $\hat{\Theta} \subset \Re^{\hat{N}}$. If $J \in \mathcal{C}_2(\hat{\Theta})$, and there exist positive constants c and C such that

$$c \,\|\, \eta \,\|^2 \leq \eta^T J''(\hat{\theta})\eta \leq C \,\|\, \eta \,\|^2 , \tag{4}$$

for all $\hat{\theta} \in \hat{\Theta}$ and for all $\eta \in \Re^{\hat{N}}$, then the gradient method beginning with any point in $\hat{\Theta}$ converges to a point $\hat{\theta}^*$. Moreover, $\hat{\theta}^*$ is the unique global minimizer of J in $\Re^{\hat{N}}$.

By the norm equivalence theorem [4], Eq. (4) is satisfied whenever $J''(\hat{\theta})$ is a *norm* on $\hat{\Theta}$ for each $\hat{\theta} \in \hat{\Theta}$. Since J'' belongs to $\mathcal{C}_2(\hat{\Theta})$, it is a symmetric matrix. It is also bounded since it is continuous over a compact set. Thus a sufficient requirement is that J'' be *positive definite* on $\hat{\Theta}$. Positive definiteness of J'' can be viewed as "positive curvature" of J at each point of $\hat{\Theta}$ which corresponds to *strict concavity* of J on $\hat{\Theta}$.

4 Taylor's Theorem

Theorem. (Taylor) Every functional $J : \Re^{\hat{N}} \mapsto \Re^1$ in $\mathcal{C}_2(\Re^{\hat{N}})$ has the representation

$$J(\hat{\theta} + \eta) = J(\hat{\theta}) + J'(\hat{\theta})\eta + \frac{1}{2}\eta^T J''(\hat{\theta} + \lambda\eta)\eta$$

for some λ between 0 and 1, where $J'(\hat{\theta})$ is the $\hat{N} \times 1$ gradient vector evaluated at $\hat{\theta} \in \Re^n$, and $J''(\hat{\theta})$ is the $\hat{N} \times \hat{N}$ Hessian matrix of J at $\hat{\theta}$, *i.e.*,

$$J'(\hat{\theta}) \triangleq \frac{\partial J(\theta)}{\partial \theta}(\hat{\theta}) \tag{5}$$

$$J''(\hat{\theta}) \triangleq \frac{\partial^2 J(\theta)}{\partial \hat{\theta}^2}(\hat{\theta}) \tag{6}$$

Proof. See Goldstein [2] p. 119. The Taylor infinite series is treated in Williamson and Crowell [7]. The present form is typically more useful for computing bounds on the error incurred by neglecting higher order terms in the Taylor expansion.

5 Newton's Method

The gradient method is based on the first-order term in the Taylor expansion for $J(\hat{\theta})$. By taking a second-order term as well and solving the quadratic minimization problem iteratively, *Newton's method* for functional minimization is obtained. Essentially, Newton's method requires the error surface to be close to *quadratic*, and its effectiveness is directly tied to the accuracy of this assumption. For most problems, the error surface can be well approximated by a quadratic form near the solution. For this reason, Newton's method tends to give very rapid ("quadratic") convergence in the last stages of iteration.

Newton's method is derived as follows: The Taylor expansion of $J(\theta)$ about $\hat{\theta}$ gives

$$J(\hat{\theta}^*) = J(\hat{\theta}) + J'(\hat{\theta})\left(\hat{\theta}^* - \hat{\theta}\right) + \frac{1}{2}\left(\hat{\theta}^* - \hat{\theta}\right)^T J''\left(\lambda\hat{\theta}^* + \overline{\lambda}\hat{\theta}\right)\left(\hat{\theta}^* - \hat{\theta}\right),$$

for some $0 \leq \lambda \leq 1$, where $\overline{\lambda} \triangleq 1 - \lambda$. It is now necessary to assume that $J''\left(\lambda\hat{\theta}^* + \overline{\lambda}\hat{\theta}\right) \approx J''(\hat{\theta})$. Differentiating with respect to $\hat{\theta}^*$, where $J(\hat{\theta}^*)$ is presumed to be minimum, this becomes

$$0 = 0 + J'(\hat{\theta}) + J''(\hat{\theta})\left(\hat{\theta}^* - \hat{\theta}\right).$$

Solving for $\hat{\theta}^*$ yields

$$\hat{\theta}^* = \hat{\theta} - [J''(\hat{\theta})]^{-1} J'(\hat{\theta}). \tag{7}$$

Applying Eq. (7) iteratively, we obtain the following.

Definition. *Newton's method* is defined by

$$\hat{\theta}_{n+1} = \hat{\theta}_n - [J''(\hat{\theta}_n)]^{-1} J'(\hat{\theta}_n), \quad n = 1, 2, \ldots, \tag{8}$$

where $\hat{\theta}_0$ is given as an initial condition.

When $J''\left(\lambda\hat{\theta}^* + \overline{\lambda}\hat{\theta}\right) = J''(\hat{\theta})$, the answer is obtained after the first iteration. In particular, when the error surface $J(\hat{\theta})$ is a *quadratic form* in $\hat{\theta}$, Newton's method produces $\hat{\theta}^*$ in one iteration, *i.e.*, $\hat{\theta}_1 = \hat{\theta}^*$ for every $\hat{\theta}_0$.

For Newton's method, there is the following general result on the existence of a critical point (*i.e.*, a point at which the gradient vanishes) within a sphere of a Banach space.

Theorem. (Kantorovich) Let $\hat{\theta}_0$ be a point in $\hat{\Theta}$ for which $[J''(\hat{\theta}_0)]^{-1}$ exists, and set

$$R_0 \triangleq \left\| [J''(\hat{\theta}_0)]^{-1} J'(\hat{\theta}_0) \right\| .$$

Let S denote the sphere $\{\hat{\theta} \in \hat{\Theta}$ such that $||\,\hat{\theta} - \hat{\theta}_0\,|| \leq 2R_0\}$. Set $C_0 = ||\,J''(\hat{\theta}_0)\,||$. If there exists a number M such that

$$\left\| J''(\hat{\theta}_1) - J''(\hat{\theta}_2) \right\| \leq \frac{M \left\| \hat{\theta}_1 - \hat{\theta}_2 \right\|}{2},$$

for $\hat{\theta}_1, \hat{\theta}_2$ in S, and such that $C_0 R_0 M \triangleq h_0 \leq 1/2$, then $J'(\hat{\theta}) = 0$ for some $\hat{\theta}$ in S, and the Newton sequence Eq. (8) converges to it. Furthermore, the rate of convergence is quadratic, satisfying

$$\left\| \hat{\theta}^* - \hat{\theta}_n \right\| \leq 2^{-n+1} (2h_0)^{2^n - 1} R_0.$$

Proof. See Goldstein [2], p. 143.

Index

References

[1] Y. Genin, "An introduction to the model reduction problem with hankel norm criterion," in *Proc. European Conf. Circuit Theory and Design, The Hague*, Aug 1981.

[2] A. A. Goldstein, *Constructive Real Analysis*, New York: Harper and Row, 1967.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations, 2nd Edition*, Baltimore: The Johns Hopkins University Press, 1989.

[4] J. M. Ortega, *Numerical Analysis*, New York: Academic Press, 1972.

[5] H. L. Royden, *Real Analysis*, New York: Macmillan, 1968.

[6] W. Rudin, *Principles of Mathematical Analysis*, New York: McGraw-Hill, 1964.

[7] R. E. Williamson, R. H. Crowell, and H. F. Trotter, *Calculus of Vector Functions*, Englewood Cliffs, NJ: Prentice-Hall, 1972.