# Gradient Based Optimization Methods

Antony Jameson*
Department of Aeronautics and Astronautics
Stanford University, Stanford, CA 94305-4035

## 1 Introduction

Consider the minimization of a function $J(x)$ where $x$ is an $n$ dimensional vector. Suppose that $J(x)$ is a smooth function with first and second derivations defined by the gradient

$$g_i(x) = \frac{\partial J}{\partial x_i}$$

and the Hessian matrix

$$A_{ij}(x) = \frac{\partial^2 J}{\partial x_i \partial x_j}$$

Generally it pays to take advantage of the smooth dependence of $J$ on $x$ by using the available information on $g$ and $A$. Suppose that there is a minimum at $x^*$ with the value $J^* = J(x^*)$. Then

$$g(x^*) = 0 \tag{1}$$

and in the neighborhood of $x^*$ $J$ can be approximated by the leading terms of a Taylor expansion as a quadratic form

$$J(x) = J^* + \frac{1}{2}(x - x^*)^T A (x - x^*) \tag{2}$$

where $A$ is evaluated at $x^*$. The minimum could be approached by a sequence of steps in the negative gradient direction

$$x_{n+1} = x_n - \beta_n g_n \tag{3}$$

where $\beta_n$ is chosen small enough to assure a decrease in $J$, or may be chosen by minimizing $J$ with a line search in the direction defined by $-g_n$. For small $\beta$ this approximates the trajectory of a dynamical system

$$\dot{x} = -\alpha g \tag{4}$$

These method are generally quite slow because the direction defined by $-g_n$ does not pass through the center of quadratic form (2 unless $x - x^*$ lies on a principal axis. A faster method is to use Newton's method to solve equation (1) and drive the gradient to zero

$$x_{n+1} = x_n - A^{-1} g_n \tag{5}$$

In complex problems it may, however, be very expensive or infeasible to determine the Hessian matrix $A$. This motivates quasi-Newton methods which recursively estimate $A$ of $A^{-1}$ from the measured changes in $g$ during the search. These methods are efficient, but if the number of variables is very large then the memory required to store the estimate of $A$ of $A^{-1}$ may become excessive. To avoid this one may use the conjugate gradient method which calculates an improved search direction by modifying the gradient to produce a vector which is conjugate to the previous search directions. This method can find the minimum of a quadratic form with $n$ line searches, but it requires the exact minimum to be found in each search direction, and it does not recover from previously introduced errors due, for example, to the fact that n general $A$ is not constant.

This note discusses some search procedures which avoid the need to store an estimate of $A$ or $A^{-1}$, and do not require exact line searches. Thus they might be suitable for problems with a very large number of variables, including the infinite dimensional case where the optimization is over the variation of a function $f(x)$.

*Thomas V. Jones Professor of Engineering, jamesonbaboon.stanford.edu

## 2 Methods based on direct estimation of the optimum

The methods are based on the idea of directly estimating the optimum $J^*$ and $x^*$ from changes on $J$ and $g$ during the search. Suppose that $J$ is sufficiently well approximated in the neighborhood of the optimum by the quadratic form (2). Then in this neighborhood

$$g(x) = A(x - x^*) \tag{6}$$

and

$$J(x) = J^* + \frac{1}{2} g^T (x - x^*) \tag{7}$$

Suppose that during a sequence of steps $X_n$ the cost and gradient are

$$J_n = J(x_n), \quad g_n = g(x_n)$$

Then we can calculate

$$y_n = J_n - \frac{1}{2} g_n^T x_n \tag{8}$$

and according to the equation (7)

$$y_n = J^* - \frac{1}{2} g_n^T x^* \tag{9}$$

Let $\hat{J}_n$ and $\hat{x}_n$ be estimates of $J^*$ and $x^*$ which are to be made at the $n^{th}$ step. Then we update $\hat{J}_n$ and $\hat{x}_n$ recursively so that

$$y_n = \hat{J}_n - \frac{1}{2} g_n^T \hat{x}_n \tag{10}$$

Define the error from substituting the previous values $\hat{J}_{n-1}$and $\hat{x}_{n-1}$ as

$$e_n = y_n - \hat{J}_{n-1} + \frac{1}{2} g_n^T \hat{x}_{n-1} \tag{11}$$

Then the general form of an update satisfying equation (10) is

$$\begin{aligned} \hat{J}_n &= \hat{J}_{n-1} + \frac{e_n v_n}{v_n - \frac{1}{2} g_n^T w_n} \\ \hat{x}_n &= \hat{x}_{n-1} + \frac{e_n w_n}{v_n - \frac{1}{2} g_n^T w_n} \end{aligned} \tag{12}$$

where $v_n$ and the vector $w_n$ may be chosen in any way such that the denominator $v_n - \frac{1}{2} g_n^T w_n$ does not vanish. Then

$$\hat{J}_n - \frac{1}{2} g_n^T \hat{x}_n \;=\; \hat{J}_{n-1} - \frac{1}{2} g_n^T \hat{x}_{n-1} + e_n \;=\; y_n$$

Alternative schemes can be derived by different rules for choosing $v_n$ and $w_n$. Also a natural choice for the steps $x_n$ is to set the new step equal to the best available estimate of the optimum

$$x_{n+1} = \hat{x}_n$$

### 2.1 Method 1

Form the augmented gradient vector $[1, -\frac{1}{2} g^T]$ and choose $v_n$ and $w_n$ so that the vector $[v_n, w_n^T]$ is orthogonal to the previous gradient vectors $[1, -\frac{1}{2} g_k^T]$, or

$$v_n - \frac{1}{2} w_n^T g_k = 0, \quad k < n$$

Suppose also that

$$\hat{J}_k - \frac{1}{2} g_k^T x_k^T = y_k, \quad k < n$$

Then

$$\begin{aligned}\hat{J}_n - \frac{1}{2}g_{n-1}^T\hat{x}_n &= \hat{J}_{n-1} + \frac{e_n v_n}{v_n - \frac{1}{2}g_n^T w_n} - \frac{1}{2}g_{n-1}^T\hat{x}_{n-1} - \frac{e_n v_n}{v_n - \frac{1}{2}g_n^T w_n} \\ &= \hat{J}_{n-1} - \frac{1}{2}g_{n-1}^T\hat{x}_{n-1} \\ &= y_{n-1}\end{aligned}$$

and by induction

$$\hat{J}_n - \frac{1}{2}g_k^T\hat{x}_n = y_k, \quad k < n$$

Now after $n+1$ evaluations

$$\hat{J}_n - \frac{1}{2}g_k^T\hat{x}_n = y_k, \quad k = 0, 1, \ldots n$$

and

$$\begin{bmatrix} 1 & -\frac{1}{2}g_{11} & \cdots & -\frac{1}{2}g_{1n} \\ 1 & -\frac{1}{2}g_{21} & \cdots & -\frac{1}{2}g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & -\frac{1}{2}g_{n1} & \cdots & -\frac{1}{2}g_{nn} \end{bmatrix} \begin{bmatrix} \hat{J}_n \\ \hat{x}_{n1} \\ \vdots \\ \hat{x}_{nn} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The same set of equations are satisfied by $J^*$ and $x^*$ if $J(x)$ is a quadratic form. Thus the minimum of a quadratic form can be found with $n+1$ evaluations of $J$ and $g$. One way to form the vectors $[v_n, w_n^T]$ would be to apply Gram Schmidt orthogonalization to the vector $[1, -\frac{1}{2}g_n^T]$. At the first step set

$$v_0 = 1, \quad w_0 = -\frac{1}{2}g_0$$

Then at the $n^{th}$ step set

$$[v_n, w_n^T] = [1, -\frac{1}{2}g_n^T] - \sum_{k=0}^{n-1} \alpha_{nk}[v_k, w_k^T]$$

where

$$\alpha_{nk} = \frac{v_k - \frac{1}{2}g_n^T w_k}{v_k^2 + w_k^T w_k}$$

This would require the storage of the previous vectors. To prevent this becoming excessive one might only force orthogonality with a limited number of previous vectors.

## 2.2 Method 2

An alternative rule is to align $w_n$ with $-g_n$. This leads to a class of updates defined by the relations

$$\begin{aligned}\hat{J}_n &= \hat{J}_{n-1} + \frac{2\alpha e_n}{\beta + \frac{1}{4}g_n^T g_n} \\ \hat{x}_n &= \hat{x}_{n-1} - \frac{\alpha e_n g_n}{\beta + \frac{1}{4}g_n^T g_n}\end{aligned}$$

where the parameters $\alpha$ and $\beta$ can be chosen to assure convergence. Define the estimation errors as

$$\tilde{J}_n = \hat{J}_n - J^*, \quad \tilde{x}_n = \hat{x}_n - x^*$$

Then

$$\begin{aligned}e_n &= y_n - \hat{J}_{n-1} + \frac{1}{2}g_{n-1}^T\hat{x}_{n-1} \\ &= \frac{1}{2}g_n^T\tilde{x}_{n-1} - \tilde{J}_{n-1} \qquad (13)\end{aligned}$$

Also set

$$\gamma = \frac{\alpha}{\beta + \frac{1}{4}g_n^T g_n}$$

Then

$$\begin{aligned}\tilde{J}_n &= \tilde{J}_{n-1} + 2\gamma e_n \\ \tilde{x}_n &= \tilde{x}_{n-1} - \gamma e_n g_n\end{aligned}$$

Therefore

$$\tilde{J}_n^2 + \tilde{x}_n^T \tilde{x}_n = \tilde{J}_{n-1}^2 + 4\gamma \tilde{J}_{n-1} e_n + 4\gamma^2 e_n^2 + \tilde{x}_{n-1}^T \tilde{x}_{n-1} - 2\gamma e_n g_n^T \tilde{x}_{n-1} + \gamma^2 e_n^2 g_n^T g_n$$

and using equation (13)

$$\tilde{J}_n^2 + \tilde{x}_n^T \tilde{x}_n - \tilde{J}_{n-1}^2 - \tilde{x}_{n-1}^T \tilde{x}_{n-1} = -\left(4\gamma - 4\gamma^2 - \gamma^2 g_n^T g_n\right) e_n^2$$

Thus the error must decrease if

$$\gamma > \gamma^2 \left(1 + \frac{1}{4} g_n^T g_n\right)$$

or

$$\alpha \frac{1 + \frac{1}{4} g_n^T g_n}{\beta + \frac{1}{4} g_n^T g_n} < 1$$

This is assured if

$$0 < \alpha < 1, \quad \beta \geq 1$$

# 3 The infinitely dimensional case

Similar ideas can be used to optimize systems where the cost $J(f)$ depends on a function $f(x)$. Define the inner product as

$$(f, g) = \int_a^b f(x) g(x) dx$$

and define the element $k = A \int$ produced by a linear operator $A$ acting on $f$ as

$$k(x) = \int_a^b A(x, x') f(x') dx'$$

Suppose the $J$ depends smoothly on $f$. Then to first order the variation $\delta J$ in $J$ which results from a variation $\delta f$ in $f$ is

$$\delta J = (g, \delta f)$$

where $g$ is the gradient. Also if $J$ reaches a minimum $J^* = J(f^*)$ at $f^*$, then the gradient is zero at the minimum. In the neighborhood of the minimum the dominant terms in the cost can therefore be represented as

$$J = J^* + \frac{1}{2} \left((f - f^*), A(f - f^*)\right) \tag{14}$$

where the operator $A$ represent the second derivative of $J$ with respect to $f$. Correspondingly the gradient can be represented near the minimum as

$$g = A(f - f^*) \tag{15}$$

Thus in this neighborhood the dominant terms in the cost can be written as

$$J = J^* + \frac{1}{2} \left(g, (f - f^*)\right) \tag{16}$$

One can now apply the techniques of Section 2 to the infinitely dimensional case. Suppose that the cost and gradient are evaluated from a sequence of trial values $f_n$, with corresponding cost $J_n$ and gradient $g_n$. We can calculate

$$y_n = J_n - \frac{1}{2}(g_n, f_n) \tag{17}$$

and according to equation (16)

$$y_n = J^* - \frac{1}{2}(g_n, f^*) \tag{18}$$

Let $\hat{J}_n$ and $\hat{f}_n$ be estimates of $J^*$ and $f^*$ at the $n^{th}$ step. These should be updated so that

$$y_n = \hat{J}_n - \frac{1}{2}(g_n, \hat{f}_n) \tag{19}$$

Define the error from the previous estimates as

$$e_n = \hat{J}_{n-1} - \frac{1}{2}(g_n, \hat{f}_{n-1})$$

Then equation (19) is satisfied by the update

$$\begin{aligned} \hat{J}_n &= \hat{J}_{n-1} + \frac{e_n v_n}{v_n - \frac{1}{2}(g_n, w_n)} \\ \hat{f}_n &= \hat{f}_{n-1} + \frac{e_n w_n}{v_n - \frac{1}{2}(g_n, w_n)} \end{aligned}$$

where $v_n$ and $w_n$ may be chosen in anyway such that the denominator $v_n - \frac{1}{2}(g_n, w_n)$ does not vanish. Then

$$\hat{J}_n - \frac{1}{2}(g_n, \hat{f}_n) \;=\; \hat{J}_{n-1} - \frac{1}{2}(g_n, \hat{f}_{n-1}) + e_n \;=\; y_n$$

Following Method 2 one can align $w_n$ with $-g_n$ and set

$$\begin{aligned} \hat{J}_n &= \hat{J}_{n-1} + \frac{2\alpha e_n}{\beta + \frac{1}{4}(g_n, g_n)} \\ \hat{f}_n &= \hat{f}_{n-1} - \frac{\alpha e_n g_n}{\beta + \frac{1}{4}(g_n, g_n)} \end{aligned}$$

Define the estimation errors as

$$\tilde{J}_n = \hat{J}_{n-1} - J^*, \quad \tilde{f}_n = \hat{f}_{n-1} - f^*$$

Then

$$\begin{aligned} e_n &= y_n - \hat{J}_{n-1} + \frac{1}{2}(g_{n-1}, \hat{f}_{n-1}) \\ &= \frac{1}{2}(g_n^T, \tilde{f}_{n-1}) - \tilde{J}_{n-1} \end{aligned} \tag{20}$$

Also set

$$\gamma = \frac{\alpha}{\beta + \frac{1}{4}(g_n, g_n)}$$

Then

$$\begin{aligned} \tilde{J}_n &= \hat{J}_{n-1} + 2\gamma e_n \\ \tilde{f}_n &= \hat{f}_{n-1} - \gamma e_n g_n \end{aligned}$$

Therefore

$$\tilde{J}_n^2 + (\tilde{f}_n, \tilde{f}_n) = \tilde{J}_{n-1}^2 + 4\gamma \tilde{J}_{n-1} e_n + 4\gamma^2 e_n^2 + (\tilde{f}_{n-1}, \tilde{f}_{n-1}) - 2\gamma e_n (g_n, \tilde{f}_{n-1}) + \gamma^2 e_n^2 (g_n, g_n)$$

and using equation (20)

$$\tilde{J}_n^2 + (\tilde{f}_n, \tilde{f}_n) - \tilde{J}_{n-1}^2 - (\tilde{f}_{n-1}, \tilde{f}_{n-1}) \;=\; -\left(4\gamma - 4\gamma^2 - \gamma^2 (g_n, g_n)\right) e_n^2$$

Thus the error must decrease if

$$\gamma > \gamma^2 \left(1 + \frac{1}{4}(g_n, g_n)\right)$$

or

$$\alpha \frac{1 + \frac{1}{4}(g_n, g_n)}{\beta + \frac{1}{4}(g_n, g_n)} < 1$$

As in the finite dimensional case, this is assured if

$$0 < \alpha < 1, \quad \beta \geq 1$$

# 4 Modified recursive procedure

The procedure proposed in sections 2 and 3 are subject to the risk that they may become ill conditioned as the optimum is approached because vectors of the form

$$[1, -\frac{1}{2}g^T]$$

will become progressively less independent as $g \to 0$. In order to circumvent this difficulty the process may recast as follows. According to equation (17) two consecutive estimates of $\hat{J}$ and $\hat{x}$ should satisfy

$$\hat{J} - \frac{1}{2}g_{n-1}^T\hat{x} = y_{n-1}$$

$$\hat{J} - \frac{1}{2}g_n^T\hat{x} = y_n$$

The first can be subtracted from the second to give

$$\delta g_n^T\hat{x} = -2\delta y_n$$

Now $n$ instances of this equation will provide sufficient information to determine the $n$-vector $\hat{x}$ provided that the changes $\delta g$ in the gradient are independent. To obtain $n$ instances will require $n+1$ evaluations of $g_n$ and $y_n$.

A recursive procedure for estimating $\hat{x}$ is as follows. Let $\hat{x}_n$ be the current estimate and let

$$e_n = 2\delta y_n + \delta g_n^T\hat{x}_n$$

be the error when this estimate is substituted in the $n^{th}$ instance. Then set

$$\hat{x}_{n+1} = \hat{x}_n - \frac{e_n}{\delta g_n^T p_n}p_n \tag{21}$$

where the step direction $p_n$ is any direction that is not orthogonal to $\delta g_n$ with the consequence that

$$\delta g_n^T\hat{x}_{n+1} = -2\delta y_n$$

Now we also require that

$$\delta g_j^T\hat{x}_{n+1} = -2\delta y_j, \quad j < n$$

This would still leave some latitude in the choice of the step direction $p_j$. However, since we expect to approach the optimum by taking a step in the negative gradient direction, provided that it is not too large, it is natural to form a set of step directions orthogonal to the changes $\delta g_j$ in the gradient from the negative gradient vectors $-g_j$. This can be achieved by a modified Gram Schmidt process. At step $n$ the new direction $p_n$ is defined recursively as follows:

$$\begin{aligned} p_n^{(1)} &= g_{n+1} \\ p_n^{(j+1)} &= p_n^{(j)} - \frac{p_n^{(j)T}\delta g_j}{p_j^T\delta g_j}, \quad j = 1, \ldots, n-1 \\ p_n &= p_n^{(n)} \end{aligned}$$

The process begins with any initial step from $x_1$ to $x_2$, giving the changes $\delta y_1$ and $\delta g_1$, and then a series of steps using the recursion formula (21), where now we move to the best available estimate, taking $x_j = \hat{x}_j, j > 1$ If the function is a quadratic form, this process will find the exact minimum after $n$ of these steps.

# 5 Acknowledgment

This document has been reproduced from "Gradient Based Optimization Methods", A. Jameson, *MAE Technical Report No. 2057*, Princeton University, 1995.