QUASI-NEWTON MODIFICATIONS TO SNOPT

KATHY JENSEN

1. Problem Statement

We address the nonlinear problem (NIP)

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & c(x) \geq 0, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable functions. Let $g(x)$ denote the gradient of $f(x)$, and $J(x)$ the mxn Jacobian matrix of $c(x)$, with rows the gradients of $c(x)$.

A point x^* is *feasible* with respect to the constraints $c(x)$ if $c(x^*) \geq 0$, and *infeasible* if $c(x^*) < 0$. The constraint $c_i(x)$ is considered *active* at x if $c_i(x) = 0$ and *inactive* if $c_i(x) > 0$.

Let $A(x)$ be the submatrix of $J(x)$ that includes only the gradients of the constraints active at x. Denote the nullspace matrix of $A(x)$ by $Z(x)$, so that the columns of $Z(x)$ form a basis for the nullspace of $A(x)$. Let the *active set* $\mathcal{A}(x)$ be the set of indices of those constraints active at x.

x^* is a *local minimizer* of NIP if it is feasible with respect to the constraints, and there exists a neighborhood of x^* such that $f(x^*) \leq f(x)$ for all feasible points x in the neighborhood. If the inequality is strict for all feasible $x \neq x^*$, then x^* is a *strong* local minimizer. Otherwise, x^* is a *weak* local minimizer.

Nonlinearly constrained problems have an abundance of diverse applications in engineering, finance, trajectory optimization, and many additional fields. The scope of scientific applications solved with nonlinearly constrained optimization software will only increase. No matter how well optimization algorithms perform, there will always be a demand for improvement as larger and more advanced mathematical models are developed.

Nonlinearly constrained problems are considerably more difficult to solve than unconstrained or linearly constrained problems, partly due to the fact that obtaining and maintaining feasibility is itself an iterative procedure. Another difficulty results from the crucial curvature

Date: April 22, 2008.

information depending upon the second derivatives of the constraints as well as those of the objective function. In the unconstrained and linearly constrained cases only the Hessian of the objective function affects the curvature.

2. Optimality Conditions

Necessary optimality conditions are guaranteed to be satisfied at a local minimizer. Once a potential solution has been computed, *sufficient optimality conditions* can verify that a point is indeed a local minimizer. Optimality conditions are invaluable in the design of efficient and robust optimization algorithms. Most widely used algorithms for solving nonlinear problems are based upon computing a point that satisfies these conditions.

In order to permit the analysis of *feasible arcs* and the determination of optimality conditions, constraints must satisfy *constraint qualifications*. Linear constraints automatically satisfy constraint qualifications. For nonlinear constraints, the theory of constraint qualifications becomes more elaborate. One possibility is that a point x is feasible and the matrix of active constraint gradients $A(x)$ has full row rank. x is then known as a *regular point* and constraint qualifications hold at x. For more details on the theory of constraint qualifications, see [4].

Assuming x^* is a regular point, first-order necessary optimality conditions state that if x^* is a local minimizer of NIP, there exist *Lagrange multipliers* λ^* such that

(1) $c(x^*) \geq 0$;
(2) $J(x^*)^T\lambda^* = g(x^*)$;
(3) $\lambda^* \geq 0$;
(4) $c(x^*)^T\lambda^* = 0$.

These first-order conditions are often referred to as the *Karush-Kuhn-Tucker* (KKT) *conditions*, and x^* is called a *KKT point*. Condition (1) says that x^* is feasible. Condition (4) is the *complementarity condition* and it says that Lagrange multipliers corresponding to inactive constraints are zero. *Strict complementarity* holds if Lagrange multipliers corresponding to active constraints are strictly positive. The *slack variables* are $s^* = c(x^*)$.

The necessary optimality conditions can alternatively be written as, there exist Lagrange multipliers μ^* such that

(1) $c(x^*) \geq 0$;
(2) $A(x^*)^T\mu^* = g(x^*)$;
(3) $\mu^* \geq 0$.

This alternate condition (2) is equivalent to $Z(x^*)^T g(x^*) = 0$, where $Z(x^*)^T g(x^*)$ is known as the *reduced gradient*.

When x^* is a regular point, the Lagrange multipliers μ^* satisfying these first-order necessary optimality conditions are unique. This can easily be shown by assuming both μ_1 and μ_2 satisfy the conditions. Then

$$A(x^*)^T \mu_1 = g(x^*) \text{ and } A(x^*)^T \mu_2 = g(x^*).$$

Subtracting these two equations we get

$$A(x^*)^T (\mu_1 - \mu_2) = 0.$$

Since the columns of $A(x^*)^T$ are linearly independent, $\mu_1 = \mu_2$.

Consider the standard Lagrangian function

$$\mathcal{L}(x, \lambda) = f(x) - \lambda^T c(x).$$

Then

$$\nabla_x \mathcal{L}(x, \lambda) = g(x) - J(x)^T \lambda,$$

and the optimality conditions say that x^* is a stationary point of the Lagrangian with respect to x when $\lambda = \lambda^*$, but not necessarily an unconstrained minimizer.

Assuming x^* is a regular point, second-order necessary optimality conditions state that if x^* is a local minimizer of NIP, then

$$Z(x^*)^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) Z(x^*) \succeq 0.$$

The matrix $Z(x^*)^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) Z(x^*)$ is known as the *reduced Hessian* of the Lagrangian.

According to sufficient optimality conditions, a regular point x^* is a strong local minimizer of NIP if there exist Lagrange multipliers λ^* such that,

(1) $c(x^*) \geq 0$;
(2) $A(x^*)^T \lambda^* = g(x^*)$;
(3) $\lambda^* > 0$;
(4) $Z(x^*)^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) Z(x^*) \succ 0$.

Condition (3) is the strict complementarity condition. See [4] for a statement of sufficient optimality conditions when Lagrange multipliers of active constraints are allowed to be zero. Even sufficient optimality conditions imply that x^* is a minimizer of the Lagrangian only on the tangent space of the active constraints, that is, the set of vectors p such that $A(x^*)p = 0$.

In section 4, we consider the nonlinear problem with equality constraints (NEP)

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & c(x) = 0. \end{aligned}$$

The definitions and theory for NEP are analogous to those for the NIP case.

Assuming x^* is a regular point, necessary optimality conditions state that if x^* is a local minimizer of NEP, there exist Lagrange multipliers λ^* such that,

(1) $c(x^*) = 0$;
(2) $J(x^*)^T\lambda^* = g(x^*)$;
(3) $Z(x^*)^T\nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)Z(x^*) \succeq 0$.

A key distinction from the inequality case is that there is no longer any restriction on the sign of the Lagrange multipliers.

Sufficient optimality conditions claim that a regular point x^* is a strong local minimizer of NEP if there exist Lagrange multipliers λ^* such that,

(1) $c(x^*) = 0$;
(2) $J(x^*)^T\lambda^* = g(x^*)$;
(3) $Z(x^*)^T\nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)Z(x^*) \succ 0$.

3. Zero or Near-Zero Lagrange Multipliers

Lagrange multipliers small in magnitude may lead to poor results. In an *active set method* for a quadratic program (QP) with inequality constraints, the quadratic is minimized over the current *working set*. The working set estimates the set of constraints active at the solution. If all the QP Lagrange multipliers are numerically positive, the original QP is solved. If any Lagrange multiplier is negative, a constraint corresponding to a negative multiplier is removed from the current working set and the process is repeated.

If a Lagrange multiplier is close to zero, computationally there may be difficulties determining the sign. If the smallest Lagrange multiplier is zero, the algorithm is at a *deadpoint*. That is, given only first-order information it is impossible to know whether or not it is beneficial to delete the corresponding constraint. Determining which subset of constraints with corresponding zero multipliers should be deleted is a combinatorial problem. [4] provides an example where deleting just one constraint with a zero multiplier does not lead to a reduction in the objective.

The example is

$$\begin{aligned} \text{minimize} \quad & -x_1x_2 \\ \text{subject to} \quad & 0 \leq x_1,x_2 \leq 10, \end{aligned}$$

with minimizer $x^* = (10, 10)$. Suppose $(0, 0)$ is the current iterate and $\{x_1 \geq 0, x_2 \geq 0\}$ is the current working set. The Lagrange multipliers corresponding to both constraints are zero. If only one constraint is removed from the working set, the objective function cannot be decreased.

4. SQP Background

Sequential quadratic programming (SQP) methods are effective in solving problems with nonlinear constraints. At each iteration, an SQP method models the original problem by a quadratic programming subproblem. The solution p_k of this subproblem is an estimate of the step from the current iterate x_k to the solution x^* of the nonlinear problem. A linesearch algorithm is implemented where the next iterate is defined as

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

with α_k a nonnegative steplength.

First we consider the equality constrained nonlinear problem. We will then extend the theory to the inequality constrained case.

4.1. SQP Method for Equality Constraints.

For the unconstrained and linearly constrained cases, optimality conditions give that the important curvature is that of the objective function. For the nonlinearly constrained case, the curvature of the Lagrangian function is key. So, we could consider forming an unconstrained quadratic model of the Lagrangian. However, as we saw, the solution is only a stationary point of the Lagrangian, but not necessarily an unconstrained minimizer. When sufficient optimality conditions hold, the solution is a minimizer only on the tangent space of active constraints. So, a quadratic model of the Lagrangian function is incomplete without constraints that restrict the space over which the minimization is occurring. Ideally, these constraints are linear.

An obvious choice is to form a quadratic model of the Lagrangian subject to the linear constraints obtained by taking the first-order approximation of the original constraints. Assume λ_k are the current estimates of the Lagrange multipliers. Let f_k, g_k, c_k and J_k be the current values of the objective, objective gradient, constraints, and Jacobian matrix, respectively. Assume J_k has full row rank. Define $\mathcal{L}_k$, $\nabla_x\mathcal{L}_k$, and $\nabla^2_{xx}\mathcal{L}_k$ to be the current values of the Lagrangian, Lagrangian

gradient, and Lagrangian Hessian, respectively. The QP subproblem obtained is:

$$\begin{aligned} \underset{p\in\mathbb{R}^n}{\text{minimize}} \quad & \mathcal{L}_k + p^T\nabla_x\mathcal{L}_k + \frac{1}{2}p^T\nabla^2_{xx}\mathcal{L}_k p \\ \text{subject to} \quad & c_k + J_k p = 0. \end{aligned}$$

Assume $\hat{p}_k$ solves this QP. Since $\hat{p}_k$ must be feasible,

$$J_k\hat{p}_k = -c_k$$

and the term

$$\hat{p}_k^T\nabla_x\mathcal{L}_k = \hat{p}_k^T(g_k - J_k^T\lambda_k)$$

in the objective function becomes

$$\hat{p}_k^T g_k + c_k^T\lambda_k.$$

Constant terms in the objective function can be discarded to obtain the equivalent QP subproblem (QP_k)

$$\begin{aligned} \underset{p\in\mathbb{R}^n}{\text{minimize}} \quad & p^T g_k + \frac{1}{2}p^T\nabla^2_{xx}\mathcal{L}_k p \\ \text{subject to} \quad & c_k + J_k p = 0. \end{aligned}$$

When the quadratic model of the Lagrangian is taken for the objective function, optimality conditions guarantee Lagrange multipliers $\hat{\lambda}_k$ such that

$$J_k^T\hat{\lambda}_k = \nabla_x\mathcal{L}_k + \nabla^2_{xx}\mathcal{L}_k\hat{p}_k.$$

Since $\nabla_x\mathcal{L}_k = g_k - J_k^T\lambda_k$, we obtain

$$J_k^T[\lambda_k + \hat{\lambda}_k] = g_k + \nabla^2_{xx}\mathcal{L}_k\hat{p}_k.$$

In the limit, as $x_k \to x^*$ and $\hat{p}_k \to 0$,

$$J(x^*)^T[\lambda_k + \hat{\lambda}_k] \approx g(x^*).$$

Compare this to the optimality condition $J(x^*)^T\lambda^* = g(x^*)$ for the nonlinearly constrained problem. $\lambda_k + \hat{\lambda}_k$ estimates λ^*, and given that $\lambda_k \to \lambda^*$, the QP Lagrange multipliers $\hat{\lambda}_k$ converge to zero. In section 3, we saw that zero or near-zero Lagrange multipliers may lead to poor results.

For the equivalent QP subproblem QP_k, optimality conditions provide Lagrange multipliers $\hat{\lambda}_k$ such that

$$J_k^T\hat{\lambda}_k = g_k + \nabla^2_{xx}\mathcal{L}_k\hat{p}_k.$$

In the limit, as $x_k \to x^*$ and $\hat{p}_k \to 0$,

$$J(x^*)^T\hat{\lambda}_k \approx g(x^*).$$

Now the QP Lagrange multipliers $\hat{\lambda}_k$ converge to the Lagrange multipliers λ^* of the original nonlinearly constrained problem. Hence, the QP subproblem QP_k is preferable, and therefore implemented in standard SQP algorithms.

An alternative derivation of the QP subproblem is based upon optimality conditions. Assume x^* is a KKT point for problem NEP and that $J(x^*)$ is full rank. Then there exist multipliers λ^* such that (x^*, λ^*) solves the nonlinear system

$$F(x, \lambda) = \begin{bmatrix} g(x) - J(x)^T \lambda \\ c(x) \end{bmatrix} = 0$$

of $n + m$ equations in $n + m$ unknowns. The Newton iterate is

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} + \begin{bmatrix} p_k \\ \Delta\lambda_k \end{bmatrix},$$

where

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} p_k \\ -\Delta\lambda_k \end{bmatrix} = \begin{bmatrix} -g_k + J_k^T \lambda_k \\ -c_k \end{bmatrix}.$$

The matrix

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L}_k & J_k^T \\ J_k & 0 \end{bmatrix}$$

is referred to as the KKT matrix. If the KKT matrix is nonsingular, the Newton step exists and is unique.

We have assumed J_k is full rank. Suppose Z_k is the nullspace matrix of J_k. Theorem 4.1 shows that when the second-order sufficient optimality condition

$$Z_k^T \nabla^2_{xx}\mathcal{L}_k Z_k \succ 0,$$

holds, then the KKT matrix is nonsingular and hence a unique Newton step exists.

Theorem 4.1. *The KKT matrix*

$$\begin{bmatrix} H & J^T \\ J & 0 \end{bmatrix}$$

is nonsingular when J has full row rank and $Z^T H Z \succ 0$, where Z is the nullspace matrix of J.

Proof. Assume

$$\begin{bmatrix} H & J^T \\ J & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0.$$

Since $Jy_1 = 0$, y_1 can be written as

$$y_1 = Zw$$

for some w.

$$HZw + J^T y_2 = 0,$$

so that

$$w^T Z^T H Z w = 0.$$

Since $Z^T HZ$ is positive definite, $w = 0$.

J being full row rank and $J^T y_2 = 0$ implies $y_2 = 0$. □

If the current iterate (x_k, λ_k) is close enough to the optimal solution (x^*, λ^*) and the sufficient optimality condition holds at (x^*, λ^*), then the sufficient optimality condition also holds at (x_k, λ_k).

The KKT system of equations gives

(1) $J_k p_k = -c_k$;
(2) $\nabla^2_{xx} \mathcal{L}_k p_k - J_k^T \Delta\lambda_k = -g_k + J_k^T \lambda_k$.

After some simple algebraic manipulations, we obtain

(1) $c_k + J_k p_k = 0$;
(2) $J_k^T \lambda_{k+1} = g_k + \nabla^2_{xx} \mathcal{L}_k p_k$.

Setting $p_k = \hat{p}_k$ and $\lambda_{k+1} = \hat{\lambda}_k$, this is the first-order optimality conditions for QP_k, which has a unique minimizer when $Z_k^T \nabla^2_{xx} \mathcal{L}_k Z_k \succ 0$.

Hence, the iterates (x_{k+1}, λ_{k+1}) can be generated either by applying Newton's method to the optimality conditions of the equality constrained nonlinear problem or by solving the quadratic subproblem QP_k. In algorithmic design, the quadratic subproblem is typically solved, and it is this approach that extends to the inequality constrained nonlinear problem.

4.2. **SQP Method for Inequality Constraints.** The SQP theory can be extended to the problem NIP by solving the QP subproblem (IQP_k)

$$\begin{aligned} \underset{p \in \mathbb{R}^n}{\text{minimize}} \quad & p^T g_k + \frac{1}{2} p^T \nabla^2_{xx} \mathcal{L}_k p \\ \text{subject to} \quad & c_k + J_k p \geq 0 \end{aligned}$$

at each iteration. The solution p_k is taken as the search direction for the original problem. As in the equality constrained case, the Lagrange multipliers $\hat{\lambda}_k$ of the QP subproblem converge to the optimal Lagrange multipliers λ^*. Hence, λ_{k+1} can be taken as $\hat{\lambda}_k$. An active set method is commonly employed to solve the quadratic subproblem.

Let A_k be the submatrix of J_k that corresponds to the constraints of IQP_k active at p_k, and $\hat{c}_k$ the corresponding constraint values, so that

$$\hat{c}_k + A_k p_k = 0.$$

Let the active set $\mathcal{A}_k$ be the set of indices of those QP constraints active at the QP solution.

Suppose the local solution x^* of NIP is a regular point at which the KKT optimality conditions are satisfied for some Lagrange multipliers λ^*. Also, suppose that strict complementarity holds at (x^*, λ^*) and that the reduced Hessian of the Lagrangian is positive definite. Then for (x_k, λ_k) sufficiently close to (x^*, λ^*), Robinson showed that the QP subproblem identifies the correct active set of the nonlinear problem at x^* [7].

Hence, the active set $\mathcal{A}_k$ of the QP subproblem predicts the optimal active set $\mathcal{A}(x^*)$ of the original nonlinear problem. When the problem is large, solving the QP subproblem from scratch at each iteration can be expensive. A substantial savings may be obtained by implementing a *warm-start* strategy where information from the previous iteration is utilized. For instance, the initial QP working set of the current subproblem can be set to the final QP active set from the previous iteration.

5. SNOPT Background

SNOPT (Sparse Nonlinear Optimizer) solves large-scale nonlinearly constrained optimization problems using a SQP algorithm. This software package was designed by Philip Gill, Walter Murray and Michael Saunders. SNOPT is equipped to handle large problems with sparse constraint gradients and functions that are expensive to compute. It was originally designed for problems with a moderate number of degrees of freedom, but has been extended to handle problems with a larger number of degrees of freedom.

We consider the nonlinear problem NIP. At each iteration of the SQP method a search direction is determined by solving the convex QP subproblem

$$\begin{aligned} \underset{p\in\mathbb{R}^n}{\text{minimize}} \quad & p^T g_k + \frac{1}{2} p^T B_k p \\ \text{subject to} \quad & c_k + J_k p \geq 0. \end{aligned}$$

Here B_k represents a positive definite approximation to the Hessian of a modified Lagrangian at x_k. B_k may not be given explicitly, but as a routine to form a matrix-vector product. In SNOPT, the subproblem is solved by SQOPT, which employs an inertia-controlling reduced Hessian active-set method.

Before defining the modified Lagrangian, we first provide the constraint linearization

$$c_L(x, x_k) = c_k + J_k(x - x_k)$$

and the departure from linearity

$$d_L(x, x_k) = c(x) - c_L(x, x_k).$$

The modified Lagrangian is defined as

$$\begin{aligned} \mathcal{L}(x, x_k, \lambda_k) = & \quad f(x) - \lambda_k^T d_L(x, x_k) \\ = & \quad \mathcal{L}(x, \lambda_k) + \lambda_k^T c_L(x, x_k). \end{aligned}$$

Since the modified Lagrangian function equals the Lagrangian plus a linear term, the Hessian of the modified Lagrangian equals that of the Lagrangian. Also,

$$\mathcal{L}(x_k, x_k, \lambda_k) = f_k$$

and

$$\nabla_x \mathcal{L}(x_k, x_k, \lambda_k) = g_k.$$

Hence, the QP subproblem is a convex quadratic model of the change in the modified Lagrangian subject to a linearization of the constraints.

We denote the solution of the current QP subproblem by $(\hat{p}_k, \hat{\lambda}_k, \hat{s}_k)$, where $\hat{s}_k$ are the slack variables. Assume the gradients of the active constraints are linearly independent. Necessary first-order optimality conditions give feasibility

$$c_k + J_k \hat{p}_k = \hat{s}_k \text{ and } \hat{s}_k \geq 0,$$

nonnegative Lagrange multipliers $\hat{\lambda}_k \geq 0$, where

$$J_k^T \hat{\lambda}_k = g_k + B_k \hat{p}_k,$$

and complementary $\hat{\lambda}_k^T \hat{s}_k = 0$.

The QP solver determines an optimal working set for the QP subproblem. This working set is an independent set of constraints that are active at the QP solution. The *working set matrix* W_k is a full-rank submatrix of J_k composed of the corresponding constraint gradients. Let Z_k be the nullspace matrix of W_k. From necessary optimality conditions, the reduced Hessian $Z_k^T B_k Z_k$ is positive semi-definite at the QP solution.

There are a couple of reasons why the change of the modified Lagrangian is modeled rather than that of the Lagrangian. First, as we will see in section 7, SNOPT sometimes utilizes the augmented modified Lagrangian

$$\mathcal{L}_A(x, x_k, \lambda_k) = f(x) - \lambda_k^T d_L(x, x_k) + \frac{1}{2} d_L(x, x_k)^T \Omega d_L(x, x_k),$$

where $\Omega = \text{diag}(\omega_i)$ and $\omega_i \geq 0$. Nonlinearity is not introduced into this augmented function. That is, any variable that appears only linearly in the original problem still appears only linearly. The SNOPT code stores nonlinear variables first and knowing that linear variables do not need to be treated as nonlinear saves memory.

A second reason for the modified Lagrangian is that, as shown in section 4, the QP subproblem Lagrange multipliers estimate the Lagrange multipliers λ^* at the solution of the original problem. For the Lagrangian the results are not so nice, and $\lambda_k + \hat{\lambda}_k$ estimate λ^*, with the QP Lagrange multipliers $\hat{\lambda}_k$ converging to zero.

Suppose s_k is the current estimate of the slack variables s^* at the solution of NIP. The search direction $\begin{pmatrix} \hat{p}_k \\ \hat{\lambda}_k - \lambda_k \\ \hat{s}_k - s_k \end{pmatrix}$ is a descent direction for an augmented Lagrangian merit function

$$M_p(x, \lambda, s) = f(x) - \lambda^T[c(x) - s] + \frac{1}{2}\sum_{i=1}^{m} \rho_i [c_i(x) - s_i]^2,$$

where $\rho_i \geq 0$ are the penalty parameters. If the QP subproblem was nonconvex, the search direction would not necessarily be a descent direction.

A merit function measures progress of an optimization algorithm. For an unconstrained problem, the objective function commonly serves as the merit function. When all the iterates remain feasible in an algorithm to solve a linearly constrained problem, the objective function is again appropriate. For a nonlinearly constrained problem, a balance between minimizing the objective and reducing infeasibility must be obtained. The augmented Lagrangian merit function $M_p(x, \lambda, s)$ balances minimizing the Lagrangian, satisfying the complementarity condition, and reducing infeasibility. Global convergence is guaranteed by performing a line search along the search direction to assure a sufficient decrease in the merit function.

6. Quasi-Newton Background

6.1. **Discrete Hessian Approximation.** When second derivatives of a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ are not available or are difficult to compute, the Hessian can be approximated at a point x_k by finite-differences of the gradient. Set the i-th column of a matrix Y to the forward-difference approximation

$$y_i = \frac{\nabla f(x_k + he_i) - \nabla f(x_k)}{h},$$

where e_i is the i-th unit vector and h is a scalar. For a symmetric Hessian approximation B_k, take

$$B_k = \frac{Y + Y^T}{2}.$$

A crucial drawback with this approximation is that it requires $n + 1$ gradient evaluations, unless there is a known sparsity pattern, in which case there may be fewer.

6.2. **Quasi-Newton Hessian Approximation.** For unconstrained optimization, a common technique for approximating the Hessian of a objective function is a quasi-Newton method. First derivatives are accumulated over a sequence of iterates $\{x_k\}$ to build up curvature information used to generate the approximate Hessian B_{k+1} at $x_{k+1} = x_k + \delta_k$. There are several desirable features of a quasi-Newton update:

(1) $B_{k+1}\delta_k = y_k$, where $y_k = g_{k+1} - g_k$.
(2) B_{k+1} is a low rank update of B_k.
(3) B_{k+1} is symmetric if B_k is.
(4) An exact line search minimizes an n degree strictly convex quadratic within n steps. If n iterates are taken, B_n is the exact Hessian.
(5) B_{k+1} is positive definite if B_k is.

Property (1) is the quasi-Newton condition, also known as the secant equation, and is required for all quasi-Newton updates. Taylor's theorem gives $\nabla^2 f(x_k)\delta_k \approx y_k$. So, the curvature of f at x_k in the direction δ_k can be approximated as $\delta_k^T \nabla^2 f(x_k)\delta_k \approx \delta_k^T y_k$. The quasi-Newton condition says this approximate curvature is exact for the local quadratic model based on B_{k+1}.

Property (3) is known as hereditary symmetry and is desirable since the Hessian of the objective function is symmetric.

Property (5) is hereditary positive-definiteness, which is desirable for two reasons. First, according to necessary optimality conditions, the Hessian is at least positive semi-definite at a minimizer. Second, a local quadratic model based on a positive definite B_k has a unique minimizer p_k, given by the Newton equation $B_k p_k = -g_k$. The search direction p_k is a descent direction for f at x_k, i.e. $p_k^T g_k = -p_k^T B_k p_k < 0$.

The unique rank-one update that satisfies (1)–(4) is the symmetric rank-one (SR1) update. This update is given as

$$B_+ = B + \frac{1}{(y - B\delta)^T \delta}(y - B\delta)(y - B\delta)^T,$$

where B_+ is the new iterate and the unsubscripted quantities are those of the current iterate.

The SR1 update has two immediate drawbacks. First, it is undefined when $(y - B\delta)^T\delta = 0$, and numerical instabilities may occur when this denominator is small. Theorem 6.1 shows that when an exact line search results in a unit step, the denominator $(y - B\delta)^T\delta$ vanishes. If $B\delta = y$, the current iterate already satisfies the quasi-Newton condition, and it makes sense to skip the update. The second major drawback is that the SR1 update does not satisfy the hereditary positive-definiteness property.

Theorem 6.1. *If an exact line search results in a unit step, then*

$$(y - B\delta)^T\delta = 0.$$

Proof. The search direction p_k is obtained by solving the Newton equation

$$B_k p_k = -g_k.$$

An exact line search gives $g_{k+1}^T p_k = 0$. A unit step gives $\delta_k = p_k$. Hence,

$$(y - B\delta)^T\delta = (y - Bp)^T p = (g_{k+1} - g_k + g_k)^T p = 0.$$

□

A higher rank update is required to satisfy properties (1)–(5). An example is the popular rank-two BFGS quasi-Newton update given by

$$B_+ = B - \frac{B\delta\delta^T B}{\delta^T B\delta} + \frac{yy^T}{y^T\delta}.$$

This update satisfies the hereditary positive-definiteness property if and only if the approximate curvature is positive. This is easy to see in one direction since if the update is positive definite, then $\delta^T y = \delta^T B_+\delta > 0$. For an outline of a proof in the other direction see [4]. Since positive-definiteness of the updates assures $\delta^T B\delta > 0$, the update is defined when positive-definiteness is maintained.

Theorem 6.2 states that the approximate curvature is positive if the step length α satisfies the strong Wolfe conditions

$$|p_k^T g_{k+1}| \leq -\eta p_k^T g_k$$

and

$$-\mu\alpha p_k^T g_k \leq f_k - f_{k+1},$$

where $0 < \mu < \eta < 1$. The strong Wolfe line search conditions guarantees a sufficient decrease in the objective function. The first condition states that the step length is not too small; the second states that the step length is not too large.

Theorem 6.2. *If the step length α satisfies the strong Wolfe conditions, then the approximate curvature is positive.*

Proof. Assume B_k is positive definite and p_k satisfies the Newton equation

$$B_k p_k = -g_k.$$

So, p_k is a descent direction, since

$$p_k^T g_k = -p_k^T B_k p_k < 0.$$

Using the first Strong Wolfe condition,

$$\delta_k^T y_k = \alpha_k p_k^T (g_{k+1} - g_k) \geq \alpha_k p_k^T (\eta g_k - g_k) = (\eta - 1)\alpha_k p_k^T g_k > 0.$$

□

When BFGS with an exact line search is applied to a quadratic with a symmetric positive definite Hessian H, induction shows that the quasi-Newton condition is satisfied by all previous search directions,

$$B_{k+1}\delta_i = y_i, \text{ for } i = 0, \ldots, k.$$

When the identity matrix is taken as the initial Hessian approximation, the search directions are conjugate with respect to H,

$$\delta_i^T H \delta_j = 0, \text{ for all } i \neq j.$$

In general, the search directions are the same as those generated by preconditioned conjugate gradient with the initial Hessian approximation as the preconditioner. When applied to an n degree strictly convex quadratic, a BFGS algorithm with an exact line search terminates within n steps. If n iterations are taken, the approximate Hessian equals the exact Hessian. As long as the updates are defined, these properties also hold for the SR1 update.

SR1 updating has even nicer results, which do not necessarily hold for BFGS. When SR1 is applied to a quadratic with a symmetric positive definite Hessian and all iterates are defined, a minimizer is obtained within n iterations. Unlike for BFGS, an exact line search is not required. If n steps are taken and the search directions are linearly independent, the exact Hessian is obtained.

Theorem 6.3 from [7] states that assuming suitable conditions, the SR1 theory extends nicely to nonlinear functions. In the limit, the approximate Hessian approaches the Hessian at the solution.

Theorem 6.3. *Assume suitable conditions. Also, assume δ_k do not converge to a linearly dependent set, and that*

$$|\delta_k^T (y_k - B_k \delta_k)| \geq r \|\delta_k\| \|y_k - B_k \delta_k\|,$$

with $r \in (0,1)$ small, so that the updates are defined and numerically stable. If $x_k \to x^*$, then

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

Proof. See [7]. □

Hence, SR1 updates tend to generate better Hessian approximations than BFGS updates. For this reason, SR1 updates are often implemented in trust-region algorithms, where indefinite Hessian approximations may even be preferred. A drawback is that when the constraint is added to restrict the step size, the trust-region subproblem may become infeasible. Handling this infeasibility is computationally expensive.

Another advantage of a SR1 update over a BFGS update is its lower rank, which leads to opportunities in limited memory quasi-Newton methods such as in SNOPT.

The quasi-Newton theory for the unconstrained case can easily be extended to the linear equality constrained problem. An approximation B_z can be generated for the reduced Hessian rather than for the Hessian so that computations are performed in a lower dimensional space. Suppose the columns of Z form a basis for the null space of the constraint matrix. Assume after the first iteration all iterates are feasible so that all steps are taken within the null space of A, i.e. a *null-space method* is implemented. If $B_z p_z = -Z^T g_k$, the search directions are given by $p_k = Z p_z$. From the quasi-Newton update for the full Hessian

$$Z^T B_+ Z = Z^T B Z - \frac{Z^T B \delta \delta^T B Z}{\delta^T B \delta} + \frac{Z^T y y^T Z}{y^T \delta}.$$

Set $\delta_z = Z^T \delta$ and $y_z = Z^T y$. If $B_z = Z^T B Z$ denotes the current reduced Hessian approximation and $\overline{B}_z$ the updated one, the quasi-Newton formula is

$$\overline{B}_z = B_z - \frac{B_z \delta_z \delta_z^T B_z}{\delta_z^T B_z \delta_z} + \frac{y_z y_z^T}{y_z^T \delta_z}.$$

An adequate line search ensures that the approximate curvature $\delta^T y = \delta_z^T y_z$ is positive at each step, so hereditary positive-definiteness holds for both the full Hessian and reduced Hessian approximations. While the reduced Hessian is at least positive semi-definite at a minimizer, the full Hessian may remain indefinite and hence may not be closely approximated.

For the linear inequality case, the approximate curvature can only be nonpositive when the line search hits a constraint. The update can then be skipped in order to maintain positive-definiteness.

7. Quasi-Newton in SNOPT

SNOPT estimates the curvature of the modified Lagrangian by using a modified BFGS quasi-Newton update to approximate the Hessian of the modified Lagrangian. Take λ_{k+1} to be the Lagrange multiplier estimates updated during the line search. The BFGS update is

$$B_+ = B - \frac{B\delta\delta^T B}{\delta^T B\delta} + \frac{yy^T}{y^T\delta},$$

where

$$\delta_k = x_{k+1} - x_k$$

and

$$y_k = \nabla\mathcal{L}(x_{k+1}, x_k, \lambda_{k+1}) - \nabla\mathcal{L}(x_k, x_k, \lambda_{k+1}).$$

Recall that the hereditary positive-definiteness property holds if and only if $y_k^T\delta_k > 0$. Unlike for the unconstrained and linear equality constrained problems, it may be impossible to obtain positive approximate curvature with any line search, even close to the solution. This makes sense since the solution is a stationary point of the Lagrangian, but not necessarily a local minimizer. Even sufficient optimality conditions only ensure a minimizer of the Lagrangian in a subspace given by the null space of the Jacobian of active constraints. So even at the solution, the Hessian of the Lagrangian may not be positive semi-definite.

Maintaining a positive definite Hessian approximation is critical so that the QP subproblem has a unique solution that is a descent direction of the augmented Lagrangian merit function. SNOPT applies an unmodified BFGS update when the approximate curvature is sufficiently positive, i.e. $y_k^T\delta_k \geq \sigma_k$ with

$$\sigma_k = \alpha_k(1-\eta)\hat{p}_k^T B_k\hat{p}_k$$

for $\eta \in (0,1)$. When there are nonlinear constraints and the approximate curvature is not sufficiently positive, SNOPT attempts to modify the BFGS update so that the approximate curvature is sufficiently positive. Otherwise, the Hessian approximation will be indefinite or ill-conditioned. If needed, two modifications are attempted.

First, SNOPT tries to modify both y_k and δ_k so that $y_k^T\delta_k$ approximates the curvature of the reduced Hessian of the Lagrangian, which from the necessary optimality conditions is positive semi-definite at the

solution. When solving the QP subproblem, let $\bar{x}_k$ be the first feasible iterate. Then decompose the search direction as

$$p_k = p_R + p_N \text{ where } p_N = \hat{x}_k - \bar{x}_k.$$

Set $z_k = x_k + \alpha_k p_R$, so that $\delta_k = \alpha_k p_N$. Redefine y_k as

$$y_k = \nabla\mathcal{L}(x_{k+1}, x_k, \lambda_{k+1}) - \nabla\mathcal{L}(z_k, x_k, \lambda_{k+1}).$$

From Taylor's theorem

$$y_k^T \delta_k \approx \alpha_k^2 p_N^T \nabla^2 \mathcal{L}(x_k, x_k, \lambda_k) p_N.$$

If the working set at $\bar{x}_k$ is the same as that at $\hat{x}_k$, then p_N is in the null space of the working set. In this case, $y_k^T \delta_k$ approximates the curvature of the reduced Hessian, which is nonnegative at the solution. Close to the solution, the working set does not change between iterations. In general, it has been observed that $W_k p_N \approx 0$, especially once changes in the working set have slowed down [3].

A disadvantage of this first modification is that it requires an extra function evaluation at z_k. However, according to [3], this modification is rarely needed more than a few times. If (x_k, λ_k) is not close to the solution, the modified $y_k^T \delta_k$ may fail to give sufficiently positive approximate curvature and a second modification is tried.

The second SNOPT modification alters only y_k by finding Δy_k so that $(y_k + \Delta y_k)^T \delta_k = \sigma_k$. This is done by resetting y_k to the difference of gradients of an augmented modified Lagrangian function

$$\mathcal{L}_A(x, x_k, \lambda_k) = f(x) - \lambda_k^T d_L(x, x_k) + \frac{1}{2} d_L(x, x_k)^T \Omega d_L(x, x_k),$$

where $\Omega = \text{diag}(\omega_i)$ and $\omega_i \geq 0$. The last term of this augmented function introduces positive curvature. The parameters ω_i are the smallest (in two-norm) that increases the approximate curvature to σ_k. Thus, the ω_i are determined by solving a linearly constrained least-squares problem. The result is

$$\Delta y_k = (J_{k+1} - J_k)^T \Omega d_L(x, x_k).$$

If neither of the attempted modifications yields a sufficiently positive approximate curvature, the update is skipped.

SNOPT provides the option to perform either a full-memory or a limited-memory quasi-Newton update. The full-memory update is appropriate when the number of nonlinear variables is small. In this case, the Hessian can be represented by a small dense matrix. Otherwise, storing the Hessian is too expensive and a limited-memory update is performed. Here only the initial Hessian approximation and the update vectors are stored. Typically, 10–20 updates are recorded and then the

diagonals are computed and stored as the initial approximation. This is enough since the important information is in the smaller dimensional reduced Hessian. Matrix-vector products involving the approximate Hessian are performed without computing the explicit approximation.

8. Proposed Modifications and Issues

The main advantages of the unconstrained BFGS update do not carry over to the nonlinearly constrained case. Since the approximate curvature may be nonpositive, the update may be undefined or numerically unstable, and hereditary positive-definiteness may not hold. The SR1 update also lacks the hereditary positive-definiteness property and may be undefined or numerically unstable. However, the SR1 update has nice features that the BFGS update lacks. For example, the SR1 update tends to generate good Hessian approximations. In the limited-memory case, only one vector and a constant need to be stored. Perhaps using the SR1 update rather than the BFGS update to approximate the Hessian of the Lagrangian is more effective.

Several issues need to be addressed when applying the SR1 update. First, the update is undefined when the denominator $(y-B\delta)^T\delta$ is zero. Numerical troubles occur when this denominator is small compared to the numerator. If $\|y - B\delta\|$ is relatively small, the current approximation almost satisfies the quasi-Newton condition, and the current update can be skipped with little effect on performance.

Otherwise, we want to check whether $y - B\delta$ and δ are almost orthogonal using

$$|(y - B\delta)^T\delta| \leq \epsilon_1\|y - B\delta\|\|\delta\|,$$

where $\epsilon_1 \in (0, 1)$ is a constant (set to 10^{-8}) [5]. In this case, the update is numerically incomputable, and the current update can be skipped or a BFGS update can be performed instead. It may be best not to perform a BFGS update since it might interfere with any nice properties of the SR1 update.

If the denominator $|(y - B\delta)^T\delta|$ is small compared to the numerator, the update matrix becomes large. If the update matrix is large compared to B_k, the n^2 bits of accumulated curvature information become overwhelmed by the n bits of new information. If the update is performed in this case, information is lost and the new approximation becomes close to a rank-one matrix. Once the approximation is essentially of rank one, it is likely to remain ill-conditioned for another $n-2$ iterations. [5] offers the check

$$\|B_{k+1} - B_k\| = \left\| \frac{(y - B\delta)(y - B\delta)^T}{(y - B\delta)^T\delta} \right\| > \gamma(1 + \|B_k\|),$$

where γ is a constant (set to 10^8). In practice, γ and ϵ_1 above are chosen to reflect the problem scaling [1]. When this check fails, the update can be skipped. Alternatively, the update can be modified so that the new approximation is better conditioned.

We want to check that B_+ is bounded and has a bounded condition number. Techniques for detecting and avoiding an ill-conditioned Hessian approximation are discussed in Walter's homework solutions. One such technique is described below.

Another consideration is that when small steps are taken, the updates should be small, especially near the solution. This may be because when small steps are taken y and δ are less reliable since they are the differences of similar numbers.

It is advantageous to maintain positive-definiteness. This guarantees that the QP subproblem will be convex and hence has a unique solution, which is a descent direction of the merit function. If positive-definiteness is maintained, the quasi-Newton method is invariant under a linear transformation of variables [5].

Suppose $B = R^T R$, where R is nonsingular. To detect indefiniteness in the updated approximation, evaluate the nonunit eigenvalue of $T = I + \frac{1}{\gamma} v v^T$, where $\gamma = (y - B\delta)^T \delta$ and $R^T v = y - B\delta$. Here $B_+ = R^T T R$ and the inertia of B_+ and T are equal. The nonunit eigenvalue of T is $\lambda = 1 + \frac{1}{\gamma} v^T v$. If λ is sufficiently positive, then to ensure positive-definiteness is not lost numerically, the Cholesky factor R is updated. The update becomes $R_+ = WR$, where $T = W^T W$ and

$$W = I + \frac{\sqrt{1 + \frac{1}{\gamma} v^T v} - 1}{v^T v} v v^T.$$

In the limited-memory scenario, only γ and v need be stored. If λ is negative, the updated approximation is indefinite. A small value of λ indicates ill-conditioning. Techniques such as those in SNOPT, i.e. changing y or δ, can be tried to increase λ. Also, modified Cholesky can be used to obtain the approximation $B_+ + E = R_+^T R_+$, where E is diagonal and R_+ nonsingular.

Another approach would be to separate the approximation of curvature into its component parts, using a SR1 update on each part. In this case, the Hessian approximation would be indefinite and many resulting issues would require attention. An advantage to this technique is that the Lagrange multipliers are not presupposed to be constant, unlike when the single update is performed.

9. Work in Progress

The theory needs to be refined, the modifications carefully implemented, and the results compared with those of the current version of SNOPT.

10. Conclusion

To be determined.

References

1. Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Mathematics of Computation **50** (1988), no. 182, 399–430.
2. Philip E. Gill, Walter Murray, and Michael A. Saunders, *User's guide for SNOPT 5.3: A fortran package for large-scale nonlinear programming.*
3. ———, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM Review **47** (2005), no. 1, 99–131.
4. Philip E. Gill, Walter Murray, and Margaret H. Wright, *Practical optimization*, Academic Press, San Diego, 1986.
5. Meredith J. Goldsmith, *Sequential quadratic programming methods based on indefinite hessian approximations*, Ph.D. thesis, Stanford University, 1999.
6. Walter Murray, *Optimization lecture notes.*
7. Jorge Nocedal and Stephen J. Wright, *Numerical optimization*, Springer, New York, 2006.
8. Walter Murray and Francisco J. Prieto, *A sequential quadratic programming algorithm using an incomplete solution of the subproblem*, SIAM J. Optim. **12** (1995), no. 4, 590–640.