Optimization

Adam Li,
Department of Applied Mathematics & Statistics,
Department of Biomedical Engineering,
Johns Hopkins University,
Baltimore, MD, 21218
ali39@jhu.edu / adam2392@gmail.com

April 23, 2021

Contents

1 Introduction

In these notes, I go over important results in optimization theory. Primarily, the focus will be on general unconstrained optimization, constrained optimization, convex optimization and then stochastic optimization. These are all big fields in themselves, so these notes serve as more of a high-level reference.

I assume working knowledge of matrix analysis and real analysis.

Currently, the first chapter deals with results in constrained optimization, namely the KKT conditions and second-order optimality conditions for general constrained problems. We will discuss algorithms in the following class of problems:

1. Linear Programming: simplex method, interior point methods, ellipsoid methods
2. Quadratic Programming: active set methods, interior point methods, gradient projection methods
3. General Optimization: penalty and augmented Lagrangian, sequential quadratic programming and interior point methods

1.1 Multivariate Calculus Short Review

Here are some useful multivariate calculus tips:

Gradient of Ax wrt x:

$$\nabla_x Ax = A$$

Taking the gradient of the quadratic form:

$$\nabla_x x^T Ax = Ax + A^T x$$

Following a post on stack-exchange, we re-state some useful facts in taking gradients of matrix multiplications with respect to vectors.

The first rule is how to take a derivative of a dot-product between two vectors:

$$\frac{\partial x^T y}{\partial x} = y$$

The second rule is the chain rule:

$$\frac{d(f(x,y))}{dx} = \frac{\partial(f(x,y))}{\partial x} + \frac{\partial y^T(x)}{\partial x}\frac{\partial f(x,y)}{\partial y}$$

where the chain rule accounts for any dependencies on x for the variable y (i.e. y might be a function of x). If y is an independent variable, then the RHS's 2nd addition becomes 0.

Let us solve for $f(x) = 1/2x^T Ax - b^T x + c$ the gradient wrt x.

$$\frac{d(b^T x)}{dx} = \frac{d(x^T b)}{dx} = b$$

and

$$\frac{d(x^T Ax)}{dx} = \frac{\partial(x^T y)}{\partial x} + \frac{d(y(x)^T)}{dx}\frac{\partial(x^T y)}{\partial y}$$

Now substituting $y = Ax$, we can arrive at the conclusion that:

$$\frac{d(x^T Ax)}{dx} = \frac{\partial(x^T y)}{\partial x} + \frac{d(y(x)^T)}{dx}\frac{\partial(x^T y)}{\partial y} = y + \frac{d(x^T A^T)}{dx}x = y + A^T x = (A + A^T)x$$

2 Constrained Optimization

Constrained optimization is now considering the minimization of problems under constraint functions. The general formulation is:

$$\min_{x\in\mathbb{R}^d} f(x) \quad s.t. \begin{cases} c_i(x) = 0, i \in \mathcal{E} \\ c_i(x) \geq 0, i \in \mathcal{I} \end{cases}$$

where $\mathcal{I}, \mathcal{E}$ are taken to be index sets for inequality constraint functions and equality constraint functions. d is the dimensionality of the data, and $f(x)$ is the objective function. Without the $c(x)$ constraint functions, this is simply an unconstrained optimization problem.

These functions are essentially **almost** assumed to be smooth. In that sense, continuously differentiable (possibly twice for second order conditions). However, note that the constraint functions $c(x)$ are not necessarily linear. They can be of general forms, but as long as they are smooth, then we can derive general conditions for optimality.

2.1 Basic Definitions

Since constraints are added, the notion of a feasible solution is different compared to unconstrained optimization problems.

Definition 2.1 (Lagrangian)**.** The Lagrangian of the general optimization problem adds a "Lagrange multiplier" that penalizes the constraint conditions:

$$L(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$$

It is important from the perspective of first-order necessary optimality conditions (i.e. the KKT conditions).

Definition 2.2 (Feasible set)**.** The feasible set Ω is the set of all points that satisfy the objective function and satisfy the constraints. Generally, we assume the objective function allows data to live in $\mathbb{R}^d$, so the feasible set is:

$$\Omega = \{x \in \mathbb{R}^d | c_i(x) = 0, i \in \mathcal{E};\ c_i(x) \geq 0, i \in \mathcal{I}\}$$

Since there are constraints, there is a notion of "active constraints", which simply tells us which constraint functions equal 0.

Definition 2.3 (Active constraint set)**.** The active set A(x) at any **feasible** point x, consists of constraints such that $c(x) = 0$.

Note: Since $c_i(x) = 0 \forall i \in \mathcal{E}$, the active set is:

$$A(x) = \mathcal{E} \cup \{i \in \mathcal{I} | c_i(x) = 0\}$$

Geometrically, we are generally very interested in the tangent vector, tangent cone, or tangent plane. Cones are geometric objects that points can arbitrarily scale by a positive value and still be within the cone.

Definition 2.4 (Tangent vector and cone)**.** The tangent vector, d, to the feasible set, Ω at a point $x \in \Omega$ if there is a feasible sequence approaching x and a sequence of positive scalars $\{t_k\} \to 0$ such that:

$$\lim_{k \to \infty} \frac{z_k - x}{t_k} = d$$

The tangent cone is the set of all tangent vectors to Ω at x and is denoted by $T_\Omega(x)$

Now, in the derivation of first-order optimality conditions, one might be interested in obtaining a feasible direction set. That is a set of directions that are going to improve the objective function. Using first-order gradient information, we can define the first-order feasible direction set.

Definition 2.5 (First order feasible direction set)**.** For a feasible point $x \in \Omega$, and active constraint set, $A(x)$, the set of linearized feasible directions is:

$$F(x) = \{d \in \mathbb{R}^d | d^T \nabla c_i(x) = 0,\ \forall i \in \mathcal{E};\ d^T \nabla c_i(x) \geq 0,\ \forall i \in \mathcal{I}\}$$

Note F(x) is a cone

2.2 Karush-Kuhn-Tucker Conditions

2.2.1 Basic Theorems and Lemmas

To prove the KKT theorem, we first prove a lemma relating the tangent cone and the first-order feasible direction set. A fundamental condition is the LICQ constraint qualification. This condition tells us that the constraints are in a sense "not redundant" because they each provide linearly "independent" information.

Definition 2.6 (Linear independence constraint qualification (LICQ))**.** Given point x and active set A(x), we say LICQ holds if the set of active constraint gradients $\{\nabla c_i(x) | i \in A(x)\}$is linearly independent.

Lemma. *If $x^* \in \Omega$ is a feasible point, then the following two statements are true:*

i) The tangent cone is a subset of the feasible direction set, $T_\Omega(x^*) \subset F(x^*)$
ii) If the LICQ condition is satisfied at x^*, then the tangent cone and feasible direction set are equivalent.

Finally, the most important setp in proving the KKT theorem is Farka's lemma, which is important because it will characterize our descent directions in optimization.

Lemma (Farka's lemma)**.** *Let cone $K = \{By + Cw | y \geq 0\} \subset \mathbb{R}^n$, where $B \in M_{n\times m}$ and $C \in M_{n\times p}$ and $y \in \mathbb{R}^m$ and $w \in \mathbb{R}^p$. K is a cone that lives in $\mathbb{R}^n$.*

Given any vector $g \in \mathbb{R}^n$, one of the two conditions holds:

i) $g \in K$ ii) there exists $d \in \mathbb{R}^n$ such that $g^T d < 0$, and $B^T d \geq 0$ and $C^T d = 0$

Note: in the second case, the vector d defines a separating hyperplane between the vector g and the cone K.

Proof. First, we will show that only one of the two is possible at a time. Assume by way of contradiction that both hold.

If $g \in K$, then there exists vecotrs $y \geq 0$ and w such that $g = By + Cw$ by definition.

If there also exists d such that $g^T d < 0$, and $B^T d \geq 0$ and $C^T d = 0$, then taking the inner product of d with g, we obtain:

$$d^t g = d^T(By + Cw) = (B^T d)^T y + (C^T d)w \geq 0$$

Now, $d^T g < 0$ by assumption, so we reach a contradiction and hence neither can hold simultaneously. Now, we show that one of the statements is true.

If $g \notin K$, then define $\hat{s} \in K$, such that $\hat{s} = \min_{s\in K} ||s - g||$. Note that K is a closed set, so $\hat{s}$ is well defined and can be the limit point of some sequence that approaches the minimum of $||s - g||$. $\alpha\hat{s} \in K$ also, by definition of a cone. We will show that $d = \hat{s} - g$ is the vector that satisfies condition ii) in the statement. Since $g \notin K$, then $d \neq 0$. Then,

$$d^T g = d^T(\hat{s} - d) = (\hat{s} - g)^T \hat{s} - d^T d$$

Earlier, we note that $\alpha = 1$ minimizes the problem $\min_\alpha ||\alpha\hat{s} - g||$, so by first-order optimality conditions, we have:

$$\frac{d}{d\alpha}||\alpha\hat{s} - g||_2^2|_{\alpha=1} = 0$$

which implies that $\hat{s}^T(\hat{s} - g) = 0$. Using this fact, we have that $d^T g = 0 - ||d||_2^2 < 0$, which is the first statement in condition ii).

Now, note thta $d^T s \geq 0$ for all $s \in K$, then the second and third statement is satisfied. □

2.2.2 First-order optimality conditions

The KKT conditions state a set of necessary conditions for any optimal solution (local, or global). They make use of gradient information for the objective and constraint functions.

Theorem (KKT First Order)**.** *Suppose x^* is a local solution and f(x) and $c_i(x)$ are continuously differentiable and that LICQ holds at x^*.*

Then there exists a Lagrange multiplier vector λ^* with dimensionality equal to $|\mathcal{E}| + |\mathcal{I}|$, such that the following conditions are satisfied at the point (x^*, λ^*):

1. *Gradient of Lagrangian evaluated at local solution is zero:* $\nabla_x L(x^*, \lambda^*) = 0$
2. *Equality constraint is satisfied:* $c_i(x^*) = 0, \ \forall i \in \mathcal{E}$
3. *Inequality constraint is satisfied:* $c_i(x^*) \geq 0, \ \forall i \in \mathcal{I}$
4. *Non-negative Lagrange multipliers for Inequality Constraints:* $\lambda_i^* \geq 0, \ \forall i \in \mathcal{I}$*, since directionality of any vector matters (because constraint is only satisfied on one side of the function!).*
5. *Complementarity Condition:* $\lambda_i^* c_i(x^*) = 0, \ \forall i \in \mathcal{E} \cup \mathcal{I}$

*Note: The λ^* is not necessarily unique if the LICQ condition does not hold, but it **will be** unique if LICQ holds.*

2.2.3 Second-order optimality conditions

The basic idea of second-order optimality conditions stems from, the fact that if we seek to minimize f(x), such that $x \in \Omega$, then:

$$w^T \nabla f(x^*) > 0 => \text{ f increases in direction of } x^* + cw$$

Therefore, first order optimality conditions tell us that we want to move in a direction opposite of the gradient to decrease the value of f(x).

However, if $w^T \nabla f(x^*) = 0$, then we will want to use second-order information.

In unconstrained optimization, note that if $\nabla^2 f(x^*) = 0$, then we might not know. In $\mathbb{R} \to \mathbb{R}$, we know that $f'(x) = 0$ implies a critical point. If $f''(x) < 0$, then it is in fact a local minima. Therefore, the sufficient second-order condition for local minima is that $f''(x) > 0$, whereas the necessary second-order condition is that $f''(x) \geq 0$.

Recall that the linearized feasible direction set at x^* is:

$$F(x^*) = \{d : d^T \nabla c_i(x^*) = 0 \ \forall i \in \mathcal{E};\ d^T \nabla c_i(x^*) \geq 0 \ \forall i \in \mathcal{I}\}$$

From the set of $F(x^*)$, we can also define the critical cone, $C(x^*, \lambda^*)$, which are a pair of points (x^*, λ^*) that satisfy the first-order KKT conditions.

Definition 2.7 (Critical Cone)**.** The critical cone $C(x^*, \lambda^*)$ is defined as follows:

$$w \in C(x^*, \lambda^*) <=> \begin{cases} \nabla c_i(x^*)^T w = 0 \ \forall i \in \mathcal{E} \\ \nabla c_i(x^*)^T w = 0 \ \forall i \in A(x^*), \lambda_i^* > 0 \\ \nabla c_i(x^*)^T w \geq 0 \ \forall i \in A(x^*), \lambda_i = 0 \end{cases} \tag{1}$$

The directions where $\lambda_i^* = 0$, then that means that direction for that constraint does not matter. Heuristically, $C(x^*, \lambda^*)$ is the set of directions where small changes to the objective remain at the boundary of the constraints.

If the gradients of the constraints at feasible point x^* is equal to 0 are linearly independent, with Lagrange multipliers, $\lambda^* > 0$, then the only directions that satisfy $w^T \nabla f(x^*) = 0$ are the ones in the critical cone. Thus the critical cone defines the directions that we need to look at because the first-order conditions tell us nothing in these directions and whether or not the objective, f, will increase or decrease. Thus the **second order condition theorems** tell us what the directions in the critical cone look like with respect to the Hessian of the Lagrangian. Note that the critical cone is a subset of the linearized feasible direction set. In addition, the critical cone is a set of "linear directions", but looking at the **curvature** of the objective function and constraint functions simultaneously.

This next theorem tells us additional necessary conditions in terms of the second-order information on the objective function. It sells us that the quadratic form of the Hessian of the Lagrangian with the critical cone directions must be non-negative in order for the point of interest to be a local solution.

Theorem (Second-order necessary optimality conditions)**.** *If x^* is a local solution satisfying LICQ and λ^* is an associated Lagrange multiplier, then taking the quadratic form (i.e. inner product of w) with the Hessian of the Lagrangian:*

$$w^T \nabla^2_{xx} L(x^*, \lambda^*) \geq 0 \ \forall w \in C(x^*, \lambda^*)$$

There are also second order **sufficient** conditions for optimality.

Theorem (Second-order sufficient optimality)**.** *Suppose that for some feasible point $x^* \in \mathbb{R}^n$, there is a Lagrange multiplier vector λ^* such that the first-order KKT conditions are satisfied. If in addition, we have:*

$$w^T \nabla^2_{xx} L(x^*, \lambda^*) w > 0$$

for all $w \neq 0 \in C(x^*, \lambda^*)$*, the critical cone. Then* x^* *is a strict local solution.*

2.2.4 Projected Hessians

As noted earlier, the Lagrange multipliers, λ^*, satisfying the KKT conditions are unique when LICQ conditions hold and strict complementarity holds. When these are unique, then the critical cone, $C(x^*, \lambda^*)$ reduces to:

$$C(x^*, \lambda^*) = Null[\nabla c_i(x^*)^T]_{i \in \mathcal{A}(x^*)} = Null A(x^*)$$

where $A(x^*)^T = [\nabla c_i(x^*)]_{i \in \mathcal{A}(x^*)}$ is the matrix with rows of active constraint gradients at x^*.

We can define the following matrix, Z, with full column rank whose columns span the critical cone space, $C(x^*, \lambda^*)$:

$$C(x^*, \lambda^*) = \{Zu | u \in \mathbb{R}^{|A(x^*)|}\}$$

That is, the critical cone consists of vectors with dimensionality equal to the number of active constraints multiplied by this Z matrix.

The condition of the 2nd-order necessary condition can be restated as:

$$u^T Z^T \nabla^2_{xx} L(x^*, \lambda^*) Z u \geq 0 \; \forall u$$

or that $Z^T \nabla^2_{xx} L(x^*, \lambda^*) Z$ is positive semidefinite by definition of PSD matrices in having non-negative quadratic form.

This matrix, Z, may actually be computed numerically, and then the conditions of the theorem may be checked by checking the eigenvalues!

First, one applies a QR factorization to the matrix of active constraint gradients, so we may obtain the null space.

2.3 Geometric Interpretation of Optimality Conditions

We can in addition to the algebraic descriptions of optimality conditions, look at it from a geometric perspective. To do so, we define various geometric objects first.

Definition 2.8 (Normal Cone)**.** The normal cone to the feasible set, Ω, at the point $x \in \Omega$ is:

$$N_\Omega(x) = \{v | v^T w \leq 0 \; \forall w \in T_\Omega(x)\}$$

where $T_\Omega(x)$ is the tangent cone at the point x.

Note: by the definition saying $v^T w \leq 0$, every normal vector, v, makes an angle of at least $\pi/2$ with every tangent vector.

Thus, we can re-write the first-order necessary condition in terms of the normal cone.

Theorem (First-order necessary condition for optimality based on normal cone)**.** *Suppose $x^* \in \Omega$ is a local minimizer of f. Then:*

$$-\nabla f(x^*) \in N_\Omega(x^*)$$

2.4 Duality in Constrained Optimization

Duality theory is a broad field that allows one to construct **alternative** formulations to a problem. Duality theory shows interesting relationships between the primal and dual problem, and in some cases the dual problem is significantly easier to solve computationally. In other cases, dual problems can be used to bound the optimal value in the primal problem.

In the primal problem, we are interested in minimizing a function of x our data. In the dual problem, we are interested in **maximizing** a function of λ, our "Lagrange multipliers".

The following is the general primal problem without equality constraints (for simplicity).

$$\min_{x \in \mathbb{R}^n} f(x) \quad s.t.\ c_i(x) \geq 0\ \forall i = 1, ..., m$$

Duality allows us to rewrite this function based on the Lagrangian:

$$L(x, \lambda) = f(x) - \lambda^T c(x)$$

such that we get the dual objective function: $q : \mathbb{R}^n \to \mathbb{R}$:

$$q(\lambda) := \inf_x L(x, \lambda)$$

It has a domain:

$$D = \{\lambda | q(\lambda) > -\infty\}$$

The computation of this infimum requires finding the **global** minimizer of the function $L(., \lambda)$ for a fixed λ, which can be arbitrarily difficult. If $L(x, \lambda)$ is convex though for a fixed λ, then we may use convex optimization to obtain a **minimizer that is global**. and then the dual optimization problem is:

$$\max_{\lambda \in \mathbb{R}^n} q(\lambda) \quad s.t.\ \lambda \geq 0$$

2.4.1 Bounding Solutions - Weak and Strong Duality

Earlier, we alluded to the fact that we could **bound** optimal solutions to the primal problem **using** solutions from the dual problem.

Theorem. *The function q is concave and its domain D is convex.*

Next, we define the term **weak-duality**, which provides a lower-bound on the optimal minimal value to the primal objective problem.

Theorem (Weak-Duality)**.** *For any $\tilde{x} \in \Omega$ feasible, and any $\tilde{\lambda} \geq 0$, we have that:*

$$q(\tilde{\lambda}) \leq f(\tilde{x})$$

2.5 Penalty Methods - Using Unconstrained with Penalty for Constrained Problems

2.6 References

Nocedal, and Wright. Springer series in operations research and financial engineering Springer, New York, NY, 2. ed. edition, (2006).

3 Linear Programming

Linear programs have by definition a **linear objective function** and **linear constraints**. The resulting feasible set is a **convex, connected polytope** set with flat, polygonal faces (think intersections of halfplanes and lines).

The linear program standard form is:

$$\min c^T x \quad s.t.\ Ax = b,\ x \geq 0$$

with $c, x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times n}$, with m constraints, and n being the dimensionality of the data points.

3.1 Duality in LP

As we saw in general duality theory, we have a way of writing optimization problems in a dual form. This dual form may, or may not be well-defined, but in the case of well-defined Linear programs, one can easily derive the dual problem.

Weak duality allows that the maximum of the dual is a lower-bound for the minimum of the primal problem. In strong duality scenarios, these two points actually coincide! Thus in strong duality, one can leverage two functional representations of the desired optimization to obtain better algorithms.

3.2 References

Nocedal, and Wright. Springer series in operations research and financial engineering Springer, New York, NY, 2. ed. edition, (2006).

4 Quadratic Programming

Now, we extend the optimization problem to a quadratic objective function and linear constraints (known as a quadratic program - QP). This type of program actually arises in many subproblems for general constrained optimziation, such as sequential quadratic programming (SQP), augmented Lagrangian methods and interior point methods.

The general QP formulation is the following:

$$\min_x q(x) = \frac{1}{2}x^T G x + x^T c \tag{2}$$

$$\text{subject to } a_i^T x = b_i, \quad i \in \mathcal{E} \tag{3}$$

$$a_i^T x \geq b_i, \quad i \in \mathcal{I} \tag{4}$$

We concern ourselves now with mainly convex QPs, where the Hessian matrix of G is positive semidefinite.

4.1 Equality-constrained QPs

First, we study the case of having only equality constraints in the QP. That is, the form of the optimization problem is:

$$\min_{x \in \mathbb{R}^n} q(x) = \frac{1}{2}x^T G x + x^T c \tag{5}$$

$$\text{subject to } Ax = b \tag{6}$$

where A is a $m \times n$ Jacobian of constraints matrix with $m \leq n$ (wide matrix), whose rows are a_i^T. Here $i \in \mathcal{E}$ and $\mathcal{I} = \phi$. $b \in \mathbb{R}^m$.

We assume that A has full row-rank (i.e. rank(A) = m), so that the constraints are consistent. First, we can analyze the first-order necessary contitions for x^* through the KKT conditions, such that that there is a vector λ^* that satsifies the following system of equations:

$$\begin{pmatrix} G & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix} \tag{7}$$

Usually, this is rewritten with $x^* = x + p$, where x is an estimate of the solution and p is the desired step. This change of variables is important because it allows us to plug in any initial estimate x, and then compute the desired offset to the optimal solution. This might be more efficient then say, directly inverting the KKT matrix.

$$K = \begin{pmatrix} G & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} -p \\ \lambda^* \end{pmatrix} = \begin{pmatrix} g \\ h \end{pmatrix} \tag{8}$$

which is also known as the **KKT matrix**. Here:

$$h = Ax - b$$

and $g = Gx + c$ and $p = x^* - x$

Note that there is a unique solution for (x^*, λ^*), when the KKT matrix is nonsingular. The next lemma gives conditions on the original constraint matrix, A, and the reduced-Hessian matrix such that the KKT matrix is non-singular. We define a matrix, Z, as a $n \times (n-m)$ matrix whose columns comprise of the null space of A.

Lemma (Sufficient conditions that KKT matrix is nonsingular)**.** *If A has full row rank, and the reduced-Hessian matrix $Z^T GZ$ is positive definite, then the KKT matrix is nonsingular.*

In addition to having a unique solution, the vector x^* is a unique global solution under the assumptions of the above lemma.

Theorem (Sufficient conditions that solution to KKT matrix problem is unique global solution)**.** *If A has full row rank, and the reduced-Hessian matrix $Z^T GZ$ is positive definite, then the vector x^* satisfying the system of equations in 8 is a global unique solution.*

Given that we have a nice theoretical properties where solving the KKT system for a nice-enough A and G matrix, then we are guaranteed a unique and global solution to our optimization problem. Now, we are interested in **efficient** methods for solving the KKT system.

4.1.1 Inertia of the KKT matrix

First, let us define the inertia of a symmetric matrix to be:

Definition 4.1 (inertia)**.** inertia(K) $= (n_+, n_-, n_0)$ where n_+, n_-, n_0 are the number of positive, negative and zero eigenvalues respectively.

Note that when $m \geq 1$, the KKT matrix is always indefinite. We can characterize the inertia of the KKT matrix:

Theorem (Inertia of the KKT matrix)**.** *Let K be the KKT matrix and suppose A has rank m. Then the inertia of K is:*

$$inertia(K) = inertia(Z^T GZ) + (m, m, 0)$$

So if $Z^T GZ$ is positive definite, then inertia(K) = (n, m, 0).

Thus this tells us that the KKT system is always indefinite when at least one constraint. To solve this system of equations, there are now a few approaches one can take.

4.1.2 Matrix Factorization Methods for Solving the KKT System

Direct methods involve matrix factorization either by: i) symmetric indefinite factorization ii) the Schur-complement method iii) Null-space method

4.1.3 Iterative Methods for Solving the KKT System

When you have a very large system, or you have a desire to implement parallelization, then an iterative approach can be taken.

4.2 Inequality Constrained QPs

When we introduce inequality constraints, now the problem becomes more complex. One can use active-set methods, or interior point methods. Before we explore methods for solving these class of problems, we review the optimality conditions for inequality-constrained QPs. First, we write the Lagrangian for this problem and then write the KKT conditions for this problem:

KKT conditions of inqequality-constrained QP

$$Gx^* + c - \sum_{i \in A(x^*)} \lambda_i^* a_i = 0 \tag{9}$$

$$a_i^T x^* = b_i \quad \text{for all } i \in A(x^*) \tag{10}$$

$$a_i^T x^* \geq b_i \quad \text{for all } i \in \mathcal{I}/\ A(x^*) \tag{11}$$

$$\lambda_i^* \geq 0 \quad \text{for all } i \in \mathcal{I} \cap A(x^*) \tag{12}$$

Similar to the equality constrained problem, if we assume that G is positive semidefinite, and a feasible point satisfies the conditions above, then we can show this proposed point is a global solution.

Theorem (Sufficient conditions for a global solution)**.** *If x^* satisfies the 9 conditions for some λ_i^* for $i \in A(x^*)$ (that is constrain i is in the active set), and G is PSD, then x^* is a global solution of the QP.*

Corollary (Uniqueness)**.** *If in addition, G is PD, then x^* is a global unique solution of the QP.*

4.3 Active set methods for convex QPs

We have seen that equality-constrained QPs are easily analyzed and solved with matrix factorization methods that essentially attempt to invert the KKT matrix to provide a solution.

We consider active-set methods here for the case of a convex QP (i.e. when the G matrix is PSD). If we knew the active set at the optimal x^*; that is $A(x^*)$, then we can find solutions using matrix factorization in the equality-constrained problem.

However, since we do not **explicitly** know which constraints are active at the optimal solution, then we must also in tandem, determine the components of this set in a solver. As a reminder, the active set methods derive from the KKT conditions that stipulate that an inequality constraint either is active, or its Lagrange multiplier is zero (and thus does not affect the objective function).

We will explore primal active-set methods, where we solve an equality-constrained QP as a subproblem based on the current **active working set**. The equality constraints are the actual equality constraints of the problem and the inequality constraints of the problem that are in the working set. Thus at each iterate, we solve the following QP subproblem:

$$\min_{p} \frac{1}{2}p^T Gp + g_k^T p \tag{13}$$

$$\text{subject to } a_i^T p = 0, \quad i \in \mathcal{W}_k \tag{14}$$

where $g_k = Gx_k + c$ and $p = x - x_k$. This subproblem will produce a solution, p_k. The p_k variable tells us at iterate k, how much to move along this direction from our x_k vector. We move in that direction a fraction of p_k that depends on where all constraints are met (including ones not in the working set). That is:

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

where $\alpha_k \in [0, 1]$ is our step size at iteration k. How might we choose α_k in a theoretically supported manner?

4.3.1 Optimal step size in QP subproblem

We can derive an explicit formula for α_k by considering what happens to the contraints that are not in the working set. If $a_i^T p_k \geq 0$ for some $i \notin W_k$, then for all $\alpha_k \geq 0$, we have the following inequality:

$$a_i^T(x_k + \alpha_k p_k) \geq a_i^T x_k \geq b_i$$

The constraint i will satisfied for non-negative α_k. Therefore, α_k is upper-bounded as:

$$\alpha_k \leq \frac{b_i - a_i^T x_k}{a_i^T p_k}$$

Since we want to decrease the objective, q, as much as possible, we want α_k as large as possible in its possible interval [0, 1], while remaining feasible.

$$\alpha_k = min(1, \min_{i \notin W_k, a_i^T p_k < 0} \frac{b_i - a_i^T x_k}{a_i^T p_k})$$

Thus, if $\alpha_k \neq 1$, then the constraints, at which the minimum is achieved, then those $i \notin W_k$ are called the blocking constraints. That is, these constraints block the program from taking a further step in the direction, p_k. We can continue these iterations, and construct a new working set, where W_{k+1} is W_k plus one of the blocking constraints from W_k. Then one can arrive at a solution, when all blocking constraints are added, and the subproblem as a solution $p = 0$.

4.3.2 Termination and Summary of the Algorithm

However, with inequality constraints, by the KKT conditions, we know that the corresponding Lagrange multipliers must be non-negative. If we find a corresponding $i \in \hat{W} \cap \mathcal{I}$, such that $\hat{\lambda}_i < 0$, then we can actually **decrease the objective** further by just removing that constraint.

The following summarizes how the active-set QP proceeds.

1. Initialize a working set, W_0
2. Compute the first-order direction to decrease the objective, p_k
3. Compute the corresponding step size to take in this direction, α_k
4. Add any blocking constraints in to form the new working set, W_k
5. Repeat the program
6. When converged, check the signage of the Lagrange multipliers of the inequality constraints
7. Remove constraints if necessary and repeat program

When we arrive at a solution, characterized by $p_k = 0$, then we have a KKT point. Furthermore, by the earlier theorem, if G is PD, then it is a unique global solution. Thus the value of the direction vector tell us our termination condition (i.e. 0 or not 0).

4.3.3 Initialization of the QP - Initial Estimate and Initial Working Set

Now, that we have examined the theoretical properties of the active-set method for convex QPs, we are interested in examining the optimal way to initialize this program. There are two things to determine: i) initial feasible point and ii) initial working set.

An initial feasible point can be given by the Phase I approach in linear programming.

To obtain the initial working set, W_0, one can simply take a linearly independent subset of the active constraints at the solution of the feasibility problem. Note that when selecting the initial working set, adding a blocking constraint and deleting a constraint cannot introduce linear dependence, so the initialization with a linearly independent constraint set ensures linear independence.

In general, this is a difficult problem.

5 Sequential Quadratic Program

These can be seen as active set methods for non-linear programming. The SQP is an effective method for nonlinearly constrained optimization, which generates steps by solving QP subproblems. We mainly consider active-set methods, which are now capable of handling significant nonlinearities in the constraints. Again, the objective program that we consider is:

$$\min_{x \in \mathbb{R}^n} f(x) \quad s.t.\ c_i(x) = 0\ i \in \mathcal{E}, \quad c_i(x) \geq 0\ i \in \mathcal{I}$$

where now, f and c are assumed to be smooth functions (i.e. continuously differentiable N times). However, we do not **assume** that c is a linear function in x. There are two approaches to such classes of problems:

1. Inequality Constrained Quadratic Program (IQP): Here, we compute an estimate of the active set and the step direction at the same time.

2. EQP: here first estimate the active set, and then solve a quadratic program with equality constraints restricted to the active set.

To understand SQP methods, we will need to:

1. develop local methods for computing the step

2. line search and trust-region methods to achieve convergence from remote starting points.

First, let us expore the local SQP with equality constraints only.

5.1 Local SQP For Equality Constraints

We consider only equality constraints first. The general approach is to form a KKT matrix and then use Newton's method for solving a system of equations. We know that the Lagrangian function for this problem is:

$$L(x, \lambda) = f(x) - \lambda^T c(x)$$

We say $A(x)^T = \big(\nabla c_1(x), \dots \nabla c_m(x)\big)$ is the Jacobian matrix of the constraints that is a $n \times m$ matrix (n dimensions and m constraints). The first-order necessary KKT conditions for optimality can be written as a system of n+m equations and n+m unknowns.:

$$F(x, \lambda) = \begin{pmatrix} \nabla f(x) - A(x)^T \lambda \\ c(x) \end{pmatrix} = 0$$

This stems from the KKT conditions. In order to solve this with nonlinearities, we can use Newton's method by:

1. Computing the Jacobian of $F(x, \lambda)$

2. Compute the Newton step for (x_k, λ_k)

3. Plug in (p_k, p_λ) to the Newton-KKT system

The Jacobian of F is:

$$F'(x, \lambda) = \begin{pmatrix} \nabla^2_{xx} L(x, \lambda) & -A(x)^T \\ A(x) & 0 \end{pmatrix}$$

and the Newon step is given by:

$$x_{k+1} = x_k + p_k, \quad \lambda_{k+1} = \lambda_k + p_\lambda$$

where (p_k, p_λ) are obtained by solving the Newon-KKT system:

$$\begin{pmatrix} \nabla^2_{xx} L(x, \lambda) & -A_k^T \\ A_k & 0 \end{pmatrix} \begin{pmatrix} p_k \\ p_\lambda \end{pmatrix} = \begin{pmatrix} -\nabla f_k + A_k^T \lambda_k \\ c_k \end{pmatrix}$$

As seen before in the linear QP, and by properties of block matrices, we know that the Newton step is well-defined with a unique solution for the step if the KKT matrix is nonsingular. This KKT matrix is nonsingular if:

1. A(x) the Jacobian matrix has full row rank (no redundant constraints)
2. The Hessian of the Lagrangian is positive definite on the tangent space of the constraints

5.2 QP SubProblem

Alternatively to the Newton method update, we can model the problem as a "sequential" QP problem, where the iterates are computed as a solution of a QP. The SQP fraamework is more practical and can be extended to inequality-constraints.

By our earlier analysis, we saw that in general solving inequality-constraint QPs involve some "guess" of the active set of constraints, and then an update. In this nonlinear inequality QP, we will see that if we can guess correctly the active set of constraints for the nonlinear problem, then we will get rapid Newton-like convergence to a local solution.

Theorem (Robinson theorem for rapid convergence of SQP for nonlinear problems)**.** *Suppose x^* is a local solution for our problem at which the KKT conditions are satisifed for some λ^*. We assume the LICQ conditions. In addition, we assume strict complementarity and second-order sufficient conditions hold at (x^*, λ^*).*

Then if (x_k, λ_k) are sufficiently close to (x^*, λ^*), there is a local solution of the QP subproblem whose active set A_k is the same as the acitve set $A(x^*)$ of the nonlinear program.

This is a qualitative theoretical result that informs us that given a **good iterate that is "close" to our optimal solution**, then our working active set will converge to the true active set of the problem at a **local solution**.

5.3 Trust-Region SQP

Trust-region methods to solve SQP are good when one does not have the Hessian matrix, $\nabla^2_{xx} L_k$, as positive definite. That is, **there is some level of control over the singularities of the Jacobian and Hessian matrices** and there is a mechanism for enforcing global convergence. A way to formulate the trust-region SQP is to add a trust-region constraint subproblem:

$$\min_p f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla_{xx}^2 L_k p \tag{15}$$

$$\text{subject to } \nabla c_i(x_k)^T p + c_i(x_k) = 0, \quad i \in \mathcal{E} \tag{16}$$

$$\nabla c_i(x_k)^T p + c_i(x_k) \geq 0, \quad i \in \mathcal{I} \tag{17}$$

$$||p|| \leq \Delta_k \quad \text{Trust region of step} \tag{18}$$

This subproblem requires us to compute a trust-region that upper-bounds the size of our direction. However, it is quite possible that this naive formulation results in an inconsistent system, where the step, p, lies outside the trust region. Hence there is no solution at that step. To tackle this issue, we review relaxation methods, penalty methods and filter methods to obtain convergent trust-region SQP methods.

5.3.1 Penalty SQP - with l_1 penalties

Here, we move the linearized constraints into the objective of the QP with a l_1 penalty term to obtain the following subproblem:

$$\min_p f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla_{xx}^2 L_k p + \mu \sum_{i \in \mathcal{E}} |c_i(x_k) + \nabla c_i(x_k)^T p| + \nu \sum_{i \in \mathcal{I}} [c_i(x_k) + \nabla c_i(x_k)^T p]^- \text{subject to } ||p||_\infty \leq \Delta_k$$

where the penalty of the inequality constraints

5.4 References

Nocedal, and Wright. Springer series in operations research and financial engineering Springer, New York, NY, 2. ed. edition, (2006).