Optimization

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

March 18, 2021

Contents

1 Introduction **1**
1.1 Multivariate Calculus Short Review 2

2 Constrained Optimization **3**
2.1 Basic Definitions 3
2.2 Karush-Kuhn-Tucker Conditions 4
2.2.1 Basic Theorems and Lemmas 4
2.2.2 First-order optimality conditions 5
2.2.3 Second-order optimality conditions 6
2.2.4 Projected Hessians 7
2.3 Geometric Interpretation of Optimality Conditions 8
2.4 Duality in Constrained Optimization 8
2.4.1 Bounding Solutions - Weak and Strong Duality 9
2.5 Penalty Methods - Using Unconstrained with Penalty for Constrained Problems 9
2.6 References 9

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:

$$\frac{d(x^TAx)}{dx} = \frac{\partial(x^Ty)}{\partial x} + \frac{d(y(x)^T)}{dx}\frac{\partial(x^Ty)}{\partial y} = y + \frac{d(x^TA^T)}{dx}x = y + A^Tx = (A + A^T)x$$

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:

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

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).