Principles of constrained optimization: inequality constraints

C8-9 of " Calculus: Real analysis and optimization"

Professor Manuel Samuelides

ISAE, University of Toulouse
Toulouse, France

October 2011

Outline

Outline

Notations

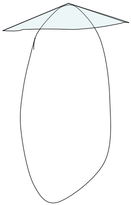
- NLP is the following optimization problem

$$\begin{cases} \min_{x\in\mathcal{E}} J(x) \\ C_E(x) = 0 \\ C_I(x) \leq 0 \end{cases}$$

  where J, C_E, C_I are smooth functions (C^2) defined on $\mathcal{E} = \mathbb{R}^n$
- The set $\Gamma = \{x \in \mathcal{E} \text{ such that } C_E(x) = 0 \text{ and } C_I(x) \leq 0$ is called the set of feasible configurations.
- If $x \in \Gamma$ for $i \in I$ a constraint inequality c_i such that $c_i(x) = 0$ is said **active** or **saturated**

Qualified constraints and admissible directions

- A direction $u \in \mathbb{R}^n$ is said **admissible** at a feasible point $x \in \Gamma$ if there exists $\epsilon > 0$ and a path $t \in [0, \epsilon[\to \gamma(t) \in \Gamma$ such that $\gamma(0) = x$ and $\gamma'(0) = u$
- since Γ is convex, the set of admissible directions at a point $x \in \Gamma$ is a convex cone noted $T_x(\Gamma)$. Note it is no more vector space since there is inequality constraints.



Properties of constraints

Proposition

If a direction u is admissible at x and if the constraint c is active at x then $\nabla c(x).u \leq 0$ with equality if c is an equality constraint.

The reciproque is not always true since some singularities may occur with inequalities constraints. So we introduce the following

Definition

The constraints are said qualified at a feasible point x if

$$[\forall i \text{ such that } c_i(x) = 0 \Rightarrow \nabla c(x)u \leq 0] \rightarrow u \in \mathcal{T}_x(\Gamma)$$

Outline

Farkas-Minkovski lemma

To prove the main theorem the biorthogonality is replaced by the similar notion for cones of bipolarity.

Definition

Let $F \subset \mathcal{E}$, then $F^o = \{u \in \mathbb{R}^n, \forall v \in F,\ (u \mid v) \geq 0\}$ is called the polar set of F. F^o is always a convex closed cone.

We have the Farkas-Minkowski lemma

Lemma

If u is in the bipolar cone of a finite set $\{c_i\}$ i.e.

$$\forall i, (c_i \mid v) \geq 0 \Rightarrow (u \mid v) \geq 0$$

then there exists $\lambda_i \geq 0$ with $u = \sum_i \lambda_i c_i$

Proof of Farkas-Minkowski Lemma

- Let K the convex cone of all positive linear combinations of c_i. If $u \notin K$, let $P_K(u)$ its projection on K. Then we have

  $$\forall v \in K, \forall t > 0, \| u - P_K(u) - tv \|^2 \geq \| u - P_K(u) \|^2$$

  from which we infer that
  $(u - P_K(u) \mid v) \leq 0 \Rightarrow P_K(u) - u \in K^o$
- We have also

  $$\| u - tP_K(u) \|^2 \geq \| u - P_K(u) \|^2 \Rightarrow (u - P_K(u) \mid P_K(u)) = 0$$

  from which we see $(u \mid P_K(u) - u) < 0$, u cannot be in the polar of K^o.

Karush-Kuhn-Tucker theorem (KKT)

Theorem

Let x^* be a local solution of NLP and suppose the constraints are qualified at x^*. Then there exists $\lambda^* = (\lambda_E^*, \lambda_I^*)$ such that

$$\begin{cases} \nabla J(x^*) + (\lambda_E^* \mid \nabla C_E(x^*)) + (\lambda_I^* \mid \nabla C_I(x^*)) = 0 \\ \forall i \in E, C_i(x^*) = 0 \\ \forall i \in I, C_i(x^*) \leq 0 \\ \forall i \in I, , \lambda_i^* \geq 0 \\ \forall i \in I, \lambda_i^*.C_i(x^*) = 0 \end{cases}$$

- The vector $\lambda^* = (\lambda_E^*, \lambda_I^*)$ is called vector of the **Kuhn-Tucker** parameters.
- If the constraint is not saturated, the associate KKT parameter is null

Proof of KKT theorem

- We follow the same lines than for Lagrange theorem for the equality constraints. So the objective gradient has to be in the polar cone of the admissible directions.
- As the constraints are qualified, the cone of admissible directions is the polar cone of the set of active constraint gradients (see previous definition).
- So the objective gradient is in the bipolar cone of active constraint gradients from which we produce Kuhn-Tucker parameters by applying Farkas-Minkovski lemma.

Obvious application

Consider NLP problem $\begin{cases} \min \frac{1}{2}(x^2+y^2) \\ x \leq a \end{cases}$ We get the Lagrangian $L(x,y,\lambda) = \frac{1}{2}(x^2+y^2) + \lambda(x-a)$.

The optimality conditions are $\begin{cases} x+\lambda = 0 \\ y = 0 \\ x \leq a \\ \lambda \geq 0 \\ \lambda(x-a) = 0 \end{cases}$

To end we have to consider wether the constraint is active or not, according to the sign of a.

Outline

Principle of trust regions algorithms

- In non-linear programming problems without or with constraints, at each step an approximate optimization problem is solved where the objective is approximated by hessian in Newton methods or by approximate inverse hessian in Quasi-Newton method (BFGS, Gauss-Newton...).
- Let v be the increment at the $k+1$-th step. The trust region method consists in adding the incremental constraint $\| v \|^2 \leq \rho^2$
- we get a new QP problem at each step $k : (x_k, \rho_k)$

$$\begin{cases} \min_v (v \mid Hv)/2 + \nabla J(x_k) v \\ C_E(x_k) + \nabla C_E(x_k) v = 0 \\ C_I(x_k) + \nabla C_I(x_k) v \leq 0 \\ \| v \|^2 \leq \rho^2 \end{cases}$$

Solution of trust regions algorithms

- To process the additional constraint we add it it in the objective introducing a lagrangian $\min_v(v \mid (H + \mu I)v)/2 + \nabla J(x_k)v$ where μ is the associate Kuhn-Tucker parameter.
- Let us discuss according to the values of ρ
  - If $\rho >> 1, \mu << 1$ the constraint is weak and we trust Newton method
  - If $\rho << 1, \mu >> 1$, the hessian is changed and the solution will be a short step in the gradient descent direction.
- So the algorithm can change continuously between a gradient descent with short step and a Newton fast converging method. How tuning ρ ?

Driving the algorithm: updating parameter

- First, it is simpler to tune μ instead of tuning ρ since in both cases the updating is empirical and incremental
- Then, the updating relies on the efficiency of previous approximations, relaxing μ if the approximation gives good result and increasing it if the approximation is poor.
- The criterion will be $R_k = \frac{J(x_{k-1}) - J(x_k)}{\frac{1}{2}(v_k | H_{k-1} v_k + \nabla J(x_{k-1}) v_k}$

So the updating rule is

Algorithm

$$\begin{cases} R_k < 0 & \Rightarrow & \mu_{k+1} = 1.75\mu_k & \textit{and reset} \\ 0 < R_k < 0.25 & \Rightarrow & \mu_{k+1} = 1.75\mu_k & \\ 0.25 < R_k < 0.75 & \Rightarrow & \mu_{k+1} = \mu_k & \\ 0.75 < R & \Rightarrow & \mu_{k+1} = 0.5\mu_k & \end{cases}$$

Outline

Notations

- Convex Programming problem (CP) stands for

$$\begin{cases} \min J(x) \\ x \in \Gamma \end{cases}$$

  with the set of following hypothesis
  1. The objective function is convex
  2. Te set of feasible configurations Γ is convex
- Practically to be convex Γ has to be defined by equality linear constraints $(c_i, i \in E)$and inequality convex constraint functions $(c_i, i \in I)$
- The unicity of the solution $(x^*, \lambda^*)$of KKT theorem is ensured by the strict convexity of J. Moreover it is generally supposed that J is α**-elliptic** with $\alpha > 0$, i.e.

$$\forall x \in \Gamma, \forall x + u \in \Gamma, \ (\nabla J(x+u) - \nabla J(x) \mid u) \geq \alpha \parallel u \parallel^2$$

KKT theorem and elliptic objective functions

Theorem

If the feasible set Γ is convex and if the objective function is elliptic, then if (x^*, λ^*) is solution of the KKT system, then x^* is the unique solution of CP (Convex Programming problem).

Proof

If x is not solution of CP, then it exists an admissible direction u such that $(\nabla J(x) \mid u) < 0$.
So ∇J is not in the bipolar of the active constraint gradient set and cannot be generated as a positive linear combination of these gradients.

Saddle-point theorem

Definition

*The **Lagrangian** of CP problem is the function*
$L(x,\lambda) = x + (\lambda \mid C(x)) = J(x) + \sum_i \lambda_i c_i(x)$.
*A couple (x^o, λ^o) is said a **saddle point** of L if*

$$\forall x \in \Gamma, \forall \lambda \in K = \mathbb{R}^{+p}, L(x^o, \lambda) \leq L(x^o, \lambda^o) \leq L(x, \lambda^o)$$

Theorem

There exists $\lambda^o \in K$ such that (x^o, λ^o) is a saddle-point of L if and only if x^o is the unique solution of CP problem.

The proof is an easy consequence of KKT theorem.

Dual problem

We can define the dual function of the objective $\mathcal{J}(\lambda) = \min_{x\in\Gamma} L(x,\lambda)$. Then for the saddle-point theorem λ^* is the solution of the dual maximization problem

$$\begin{cases} \max \mathcal{J}(\lambda) \\ \lambda \in K \end{cases}$$

The dual problem may be simpler to solve if we get the dual function since the constraints are **bound** constraints and can be solved using the projected gradient algorithm.

Outline

1. Inequality constraints
   - Notations and definitions
   - Karush-Kuhn-Tucker theorem
   - Trust region algorithm
2. Convex programming and duality
   - Saddle point theorem
3. Algorithmics
   - Projected gradient
   - Usawa algorithm

Principle of projected gradient

- The idea is to minimize the objective according to a gradient descent and then to come back to the feasible configuration set Γ by a projection $P_G a$
- It is applied when Γ is convex
- This algorithm is useful when the constraints are bound constraints, (which is often the case in dual problems). In that case, P_K is easy to compute.

Algorithm

1. *Compute a descent algorithm* $y_{k+1} = x_k - \rho \nabla J x_k$
2. *Project on the feasible set* Γ*:* $x_{k+1} = P_\Gamma(y_{k+1})$

Projection on a convex set

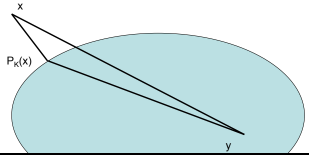
The proof of convergence relies on the following properties of the projection on a convex set:

Lemma

The projection P_Γ onto a convex set checks
$\forall x \in \mathcal{E}, \forall y \in \Gamma, (x - P_\Gamma(x) \mid y - P_\Gamma(x)) \leq 0$

Lemma

P_Γ is a contraction



Proof of convergence of projected gradient for QP

- Let x^* be the solution, then we notice that $P_\Gamma(x^* - \rho\nabla J(x^*)) = x^*$ from the extremality condition. Then x^* is a fixed point of the algorithm.
- Then we have

$$\| x_{k+1} - x^* \| \leq \| [x_k - \rho(Ax_k - b] - [x^* - \rho(Ax^* - b)] \|$$

$$\| x_{k+1} - x^* \| \leq \| (I - \rho A)(x_k - x^*) \|$$

- The algorithm converges linearly for $\rho < \frac{2}{\|A\|}$

Outline

Description of Usawa algorithm

In case J and C are convex, the convex programming problem (CP) $\begin{cases} \min J(x) \\ C(x) \leq 0 \end{cases}$ can be solved by a primal-dual method which alternatively updates the primal variables and the dual ones:

Algorithm

1. $x_{k+1} = \arg\min_x L(x, \lambda_k)$
2. $\lambda_{k+1} = P_{\{\lambda \geq 0\}}[\lambda_k + \rho C(x_{k+1})]$ *with* ρ *small enough*

NLP solution is fixed by Usawa algorithm

- To prove the convergence of Usawa algorithm strong convexity assumptions have to be set (ellipticity of J, smoothness of C will tune the step ρ.
- It is very easy to check that the solution is the fixed point of Usawa algorithm.
- Then the convergence amounts to the convergence of projected gradient using the ellipticity of the objective function and the Lipschitz continuity of the constraint.

Summary

- We generalize the Lagrange theorem for inequality constraints withe KKT theorem.
- The KT parameters for active inequality constraints are positive.
- It is a problem to choose active constraints before computing solution.
- The constraints are more often convex than for non linear equality constrained problems. Then the saddle-point theorem may be applied to transform the problem and solve the dual problem first.
- In the dual problem constraints are only bound constraints and the projected gradient method may be applied.
- Usawa's algorithm is a practical application of the saddle-point theorem using projected gradient algorithm.

For Further Reading I

- J.F.. Bonnans, J.C.. Gilbert, C. Lemaréchal, C.A. Sagastizabal.
  Numerical optimization, theoretical and practical aspects ,
  Part I, Springer (2009).
- R.Fletcher.
  Practical methods of optimization ,
  Wiley (1987).
- M.Samuelides.
  Optimisation numérique ,
  Polycopié SUPAERO (1998).