


Trust-Region Methods

Introduction

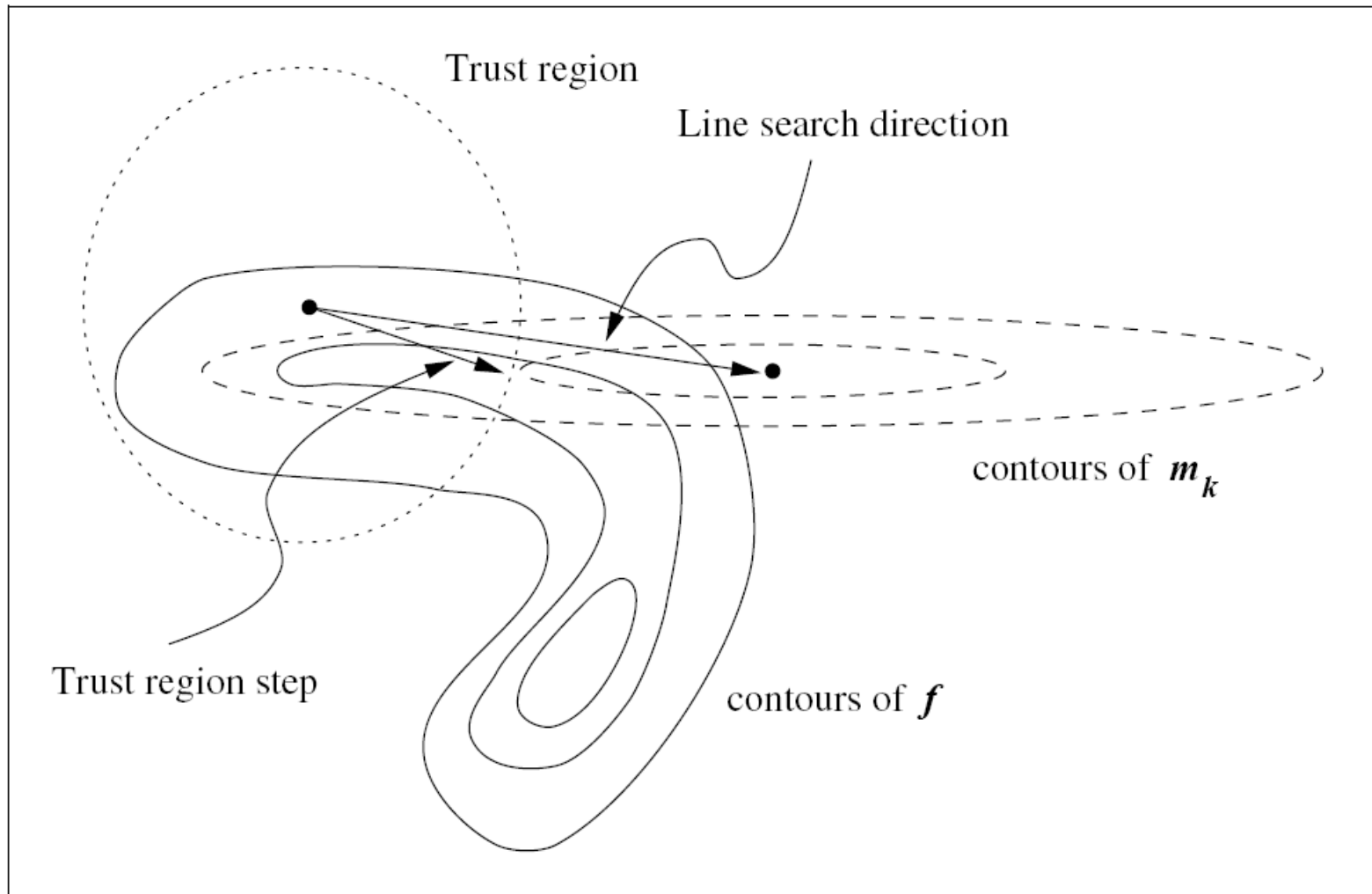
Line search methods and trust-region methods both generate steps with the help of a quadratic model of the objective function, but they use this model in different ways.

Line search — generate a search direction

Trust-region — define a region around the current iterate

- 
- Choose the direction and length of the step simultaneously
 - In general, **the step direction changes** whenever the size of the trust region is altered.

Trust-region and line search steps



Introduction

We will assume that the first two terms of the quadratic model functions m_k at each iterate x_k are identical to the first two terms of the Taylor-series expansion of f around x_k .

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T \underbrace{B_k}_{\text{some symmetric matrix}} p$$

Taylor-series: $f(x_k + p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla^2 f(x_k + tp) p$

$O(\|p\|^2)$

So the approximation error is small when p is small.

Introduction

If $B_k = \nabla^2 f(x_k) \Rightarrow$ the approximation error is $O(\|p\|^3)$

— The algorithm based on setting $B_k = \nabla^2 f(x_k)$ is called the **Trust-region Newton Method**.

For the general case, we only assume that B_k is *symmetric* and *uniformly bounded* in the index k .

To obtain each step, we seek a solution of the subproblem

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \quad \text{s.t. } \|p\| \leq \Delta_k \quad (1)$$

↓
 $\|\cdot\|_2 \rightarrow$ ball

Introduction

When B_k is positive definite and $\|B_k^{-1}\nabla f_k\| \leq \Delta_k$, the solution of the subproblem is easy to identify.

$$p_k^B = -B_k^{-1}\nabla f_k \quad \text{— Full Step}$$

- In other cases, the solution of the subproblem is not so obvious.
- In any case, we need only an *approximate* solution to obtain convergence and good practical behavior.

Outline of the algorithm

We base the choice of trust-region radius Δ_k on the agreement between the model function m_k and the objective function f at previous iterations.

Given a step p_k we define the ratio

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{\boxed{m_k(0) - m_k(p_k)}} \begin{array}{l} \rightarrow \text{actual reduction} \\ \rightarrow \text{predicted reduction} \end{array}$$

always ≥ 0

if ρ_k is negative, the new objective value $f(x_k + p_k)$ is greater than the current value $f(x_k)$, so the step must be rejected.

Outline of the algorithm

On the other hand, ρ_k is positive, then

Close to 1 — there is good agreement over this step, so it is safe to expand the trust region for the next iteration.

Not close to 1 — do not alter the trust region, but if it is close to zero or negative, we shrink the trust region.

Algorithm (Trust Region).

Given $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$, and $\eta \in [0, \frac{1}{4})$:

for $k = 0, 1, 2, \dots$

Obtain p_k by (approximately) solving (1) and evaluate ρ_k ;

an overall bound

Outline of the algorithm

if $\rho_k < \frac{1}{4}$

$$\Delta_{k+1} = \frac{1}{4} \|p_k\| \longrightarrow \text{shrink}$$

else

if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$ reach the boundary

$$\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta}) \longrightarrow \text{expand}$$

else

$$\Delta_{k+1} = \Delta_k; \longrightarrow \text{do not alter}$$

if $\rho_k > \eta$

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

else

$$x_{k+1} = x_k;$$

end (for).

The Cauchy point

It is enough for global convergence purposes to find an approximate solution p_k that lies within the trust region and gives a sufficient reduction in the model.



can be quantified in terms of the Cauchy point

Cauchy Point Calculation

Find the vector p_k^s that solves a linear version of (1), that is,

$$p_k^s = \arg \min_{p \in R^n} f_k + \nabla f_k^T p \quad \text{s.t.} \quad \|p\| \leq \Delta_k$$

Calculate the scalar $\tau_k > 0$ such that

$$\tau_k = \arg \min_{\tau > 0} m_k(\tau p_k^s) \quad \text{s.t.} \quad \|\tau p_k^s\| \leq \Delta_k$$

$$p_k^c = \tau_k p_k^s$$

The Cauchy point

$$p_k^s = -\frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k$$

To obtain τ_k explicitly, we consider the cases of $\nabla f_k^T B_k \nabla f_k \leq 0$ and $\nabla f_k^T B_k \nabla f_k > 0$ separately.

○ $\nabla f_k^T B_k \nabla f_k \leq 0$

$m_k(\tau p_k^s)$ decreases monotonically with τ whenever $\nabla f_k \neq 0$

\Rightarrow the largest value $\tau_k = 1$

○ $\nabla f_k^T B_k \nabla f_k > 0$

$m_k(\tau p_k^s)$ is a convex quadratic in τ $\begin{cases} \|\nabla f_k\|^3 / (\Delta_k \nabla f_k^T B_k \nabla f_k) \\ 1 \end{cases}$

The Cauchy point

In summary, we have

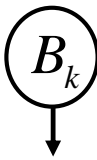
$$p_k^C = -\tau_k \frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k$$

where

$$\tau_k = \begin{cases} 1 & \text{if } \nabla f_k^T B_k \nabla f_k \leq 0; \\ \min\left(\|\nabla f_k\|^3 / (\Delta_k \nabla f_k^T B_k \nabla f_k), 1\right) & \text{otherwise.} \end{cases}$$

The Cauchy step p_k^C is inexpensive to calculate and is of crucial importance in deciding if an approximate solution of the trust-region subproblem is acceptable.

Improving on the Cauchy point

- ◆ The Cauchy point does not depend very strongly on the matrix B_k , which is used only in the calculation of the step length.
- ◆ Rapid convergence (superlinear, for instance) can be expected only if B_k plays a role in determining the direction of the step as well as its length.
- ◆ The improvement strategy is often designed so that the full step $p_k^B = -B_k^{-1} \nabla f_k$ is chosen whenever B_k is positive definite and $\|p_k^B\| \leq \Delta_k$.
the exact Hessian or a quasi-Newton approximation can be expected to yield superlinear convergence

The dogleg method

When B is positive definite, we have already noted that the unconstrained minimizer of m is the full step $p^B = -B^{-1}g$.

$$\min_{p \in \mathcal{R}^n} m(p) \stackrel{\text{def}}{=} f + g^T p + \frac{1}{2} p^T B p \quad \text{s.t. } \|p\| \leq \Delta$$

$$p^*(\Delta) = p^B, \quad \text{when } \Delta \geq \|p^B\| \quad \text{Feasible}$$

When Δ is tiny, the restriction $\|p\| \leq \Delta$ ensures that the quadratic term in m has little effect on the solution.

$$p^*(\Delta) \approx -\Delta \frac{g}{\|g\|}, \quad \text{when } \Delta \text{ is small.}$$

The dogleg method

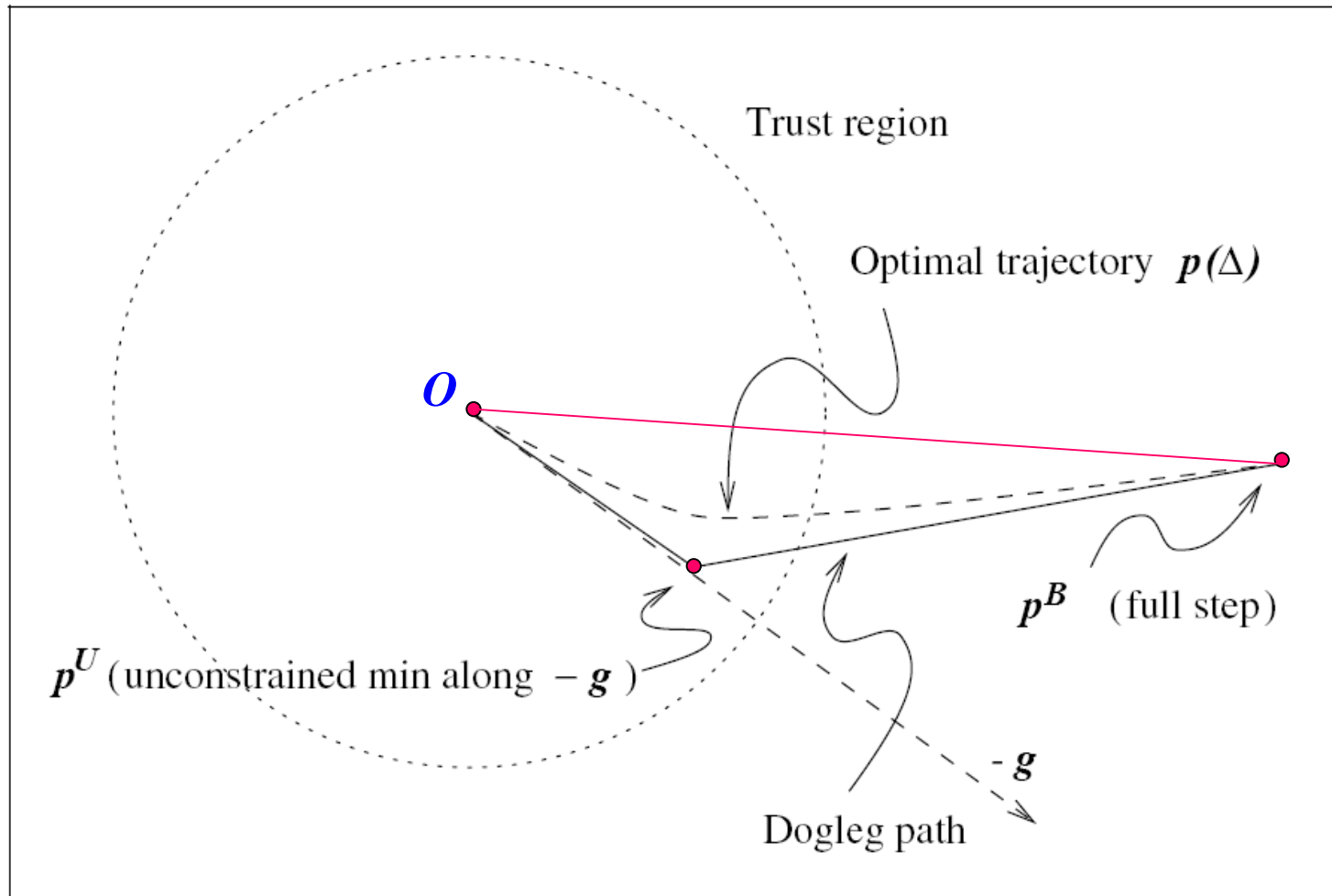
For intermediate values of Δ , the solution $p^*(\Delta)$ typically follows a curved trajectory like the one in the following figure. ►

The *dogleg method* finds an approximate solution by replacing the curved trajectory for $p^*(\Delta)$ with a path consisting of two line segments.

—— **The first line segment:** it runs from the origin to the unconstrained minimizer along the steepest descent direction

$$p^U = -\frac{g^T g}{g^T B g} g$$

Exact trajectory and dogleg approx.



The dogleg method

—— The second line segment: it runs from p^U to p^B .

Formally, we denote this trajectory by $\tilde{p}(\tau)$ for $\tau \in [0, 2]$, where

$$\tilde{p}(\tau) = \begin{cases} \tau p^U, & 0 \leq \tau \leq 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2. \end{cases}$$

The dogleg method chooses p to minimize the model m along this path, subject to the trust-region bound.

In fact, it is not even necessary to carry out a search because the dogleg path intersects the trust-region boundary at most once and the intersection point can be computed analytically.

Lemma

Let B be positive definite. Then

- (i) $\|\tilde{p}(\tau)\|$ is an increasing function of τ , and
- (ii) $m(\tilde{p}(\tau))$ is a decreasing function of τ .

Proof. It is easy to show that (i) and (ii) both hold for $\tau \in [0, 1]$, so we restrict our attention to the case of $\tau \in [1, 2]$.

$$\begin{aligned}\text{For (i), define } h(\alpha) &= \frac{1}{2} \|\tilde{p}(1 + \alpha)\|^2 \\ &= \frac{1}{2} \|p^U + \alpha(p^B - p^U)\|^2 \\ &= \frac{1}{2} \|p^U\|^2 + \alpha p^{U^T} (p^B - p^U) + \frac{1}{2} \alpha^2 \|p^B - p^U\|^2.\end{aligned}$$

Lemma (cont'd)

Our result is proved if we can show that $h'(\alpha) \geq 0$ for $\alpha \in (0,1)$.

$$\begin{aligned} h'(\alpha) &= -p^{U^T} (p^U - p^B) + \alpha \|p^U - p^B\|^2 \\ &\geq -p^{U^T} (p^U - p^B) \\ &= \frac{g^T g}{g^T B g} g^T \left(-\frac{g^T g}{g^T B g} g + B^{-1} g \right) \\ &= g^T g \frac{g^T B^{-1} g}{g^T B g} \left[1 - \frac{(g^T g)^2}{(g^T B g)(g^T B^{-1} g)} \right] \\ &\geq 0, \end{aligned}$$

Lemma (cont'd)

For (ii), we define $\tilde{h}(\alpha) = m(\tilde{p}(1+\alpha))$ and show that $\tilde{h}'(\alpha) \leq 0$ for $\alpha \in (0,1)$.

$$\begin{aligned}\tilde{h}'(\alpha) &= (p^B - p^U)^T (g + Bp^U) + \alpha (p^B - p^U)^T B (p^B - p^U) \\ &\leq (p^B - p^U)^T (g + Bp^U + B(p^B - p^U)) \\ &= (p^B - p^U)^T (g + Bp^B) = 0\end{aligned}$$

It follows from this lemma that the path $\tilde{p}(\tau)$ intersects the trust-region boundary $\|p\| = \Delta$ at exactly one point if $\|p^B\| \geq \Delta$, and nowhere otherwise.

$$\|p^U + (\tau - 1)(p^B - p^U)\|^2 = \Delta^2$$

Global convergence

Reduction obtained by the CP:

$$m_k(0) - m_k(p_k^c) \geq \frac{1}{2} \|\nabla f_k\| \min \left(\Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right)$$

by the dogleg method:

$$m_k(0) - m_k(p_k) \geq c_1 \|\nabla f_k\| \min \left(\Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right) \quad (2)$$

$$c_1 \in (0, 1]$$

Usually, $m_k(p_k) \leq m_k(p_k^c)$

Global convergence

General Assumptions:

$\|B_k\| \leq \beta$ for some constant β — uniformly bounded in norm

f is continuously differentiable and bounded below

all approximate solutions of (1) satisfy the inequalities (2) and

$$\|p_k\| \leq \gamma \Delta_k, \quad \text{for some constant } \gamma \geq 1$$

$$\eta = 0$$

$$\liminf_{k \rightarrow \infty} \|\nabla f_k\| = 0$$

$$\eta \in \left(0, \frac{1}{4}\right)$$

$$\lim_{k \rightarrow \infty} \nabla f_k = 0$$



Chapter 4

Fundamentals of Constrained Optimization

Introduction

A general formulation is

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{cases} c_i(x) = 0, & i \in E, \text{ — equality constraints} \\ c_i(x) \geq 0, & i \in I, \text{ — inequality constraints} \end{cases}$$

where f and the functions c_i are all smooth.

e.g., twice continuously differentiable, etc.

— *feasible set*:

$$\Omega = \{x \mid c_i(x) = 0, i \in E; c_i(x) \geq 0, i \in I\}$$



The compact form: $\min_{x \in \Omega} f(x)$

Review

Recall that for the unconstrained optimization problem, we characterized solution points x^* in the following way:

Necessary conditions: Local minima of unconstrained problems have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ positive semidefinite.

Sufficient conditions: Any point x^* at which $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite is a strict local minimizer of f .

Our aim in this part is to derive similar conditions to characterize the solutions of constrained optimization problems.

Local and global solutions

We have seen already that global solutions are difficult to find even when there are no constraints. Then how about the situation when we add constraints?

Example1: $\min_{x \in \mathbb{R}^n} \|x\|_2^2$, subject to $\|x\|_2^2 \geq 1$

→ Any vector x with $\|x\|_2 = 1$ solves the problem.

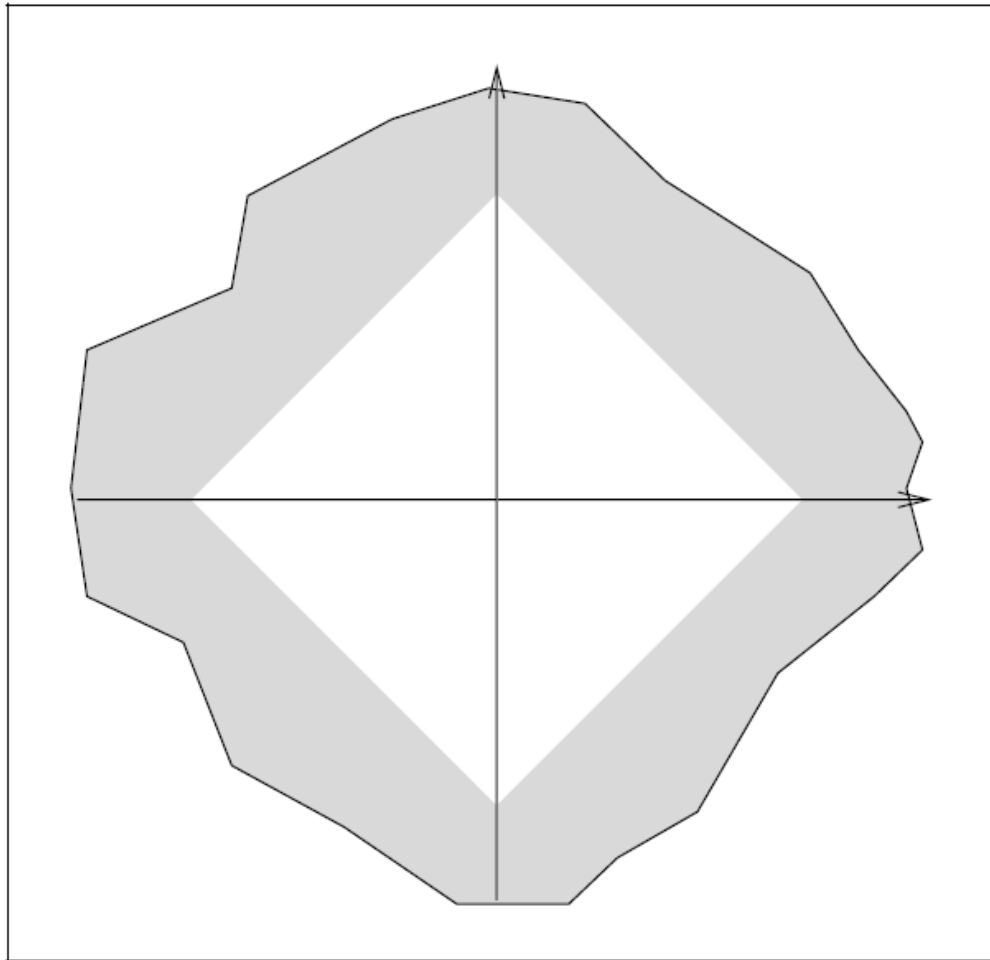
infinitely many local minima

Example2:

$\min (x_2 + 100)^2 + 0.01x_1^2$, subject to $x_2 - \cos x_1 \geq 0$

$(x_1, x_2) = (k\pi, -1)$, for $k = \pm 1, \pm 3, \pm 5, \dots$

Smoothness



Single constraint:

$$\|x\|_1 = |x_1| + |x_2| \leq 1$$



A set of smooth constraints:

$$x_1 + x_2 \leq 1,$$

$$x_1 - x_2 \leq 1,$$

$$-x_1 + x_2 \leq 1,$$

$$-x_1 - x_2 \leq 1.$$

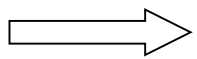
Example 1

Definition:

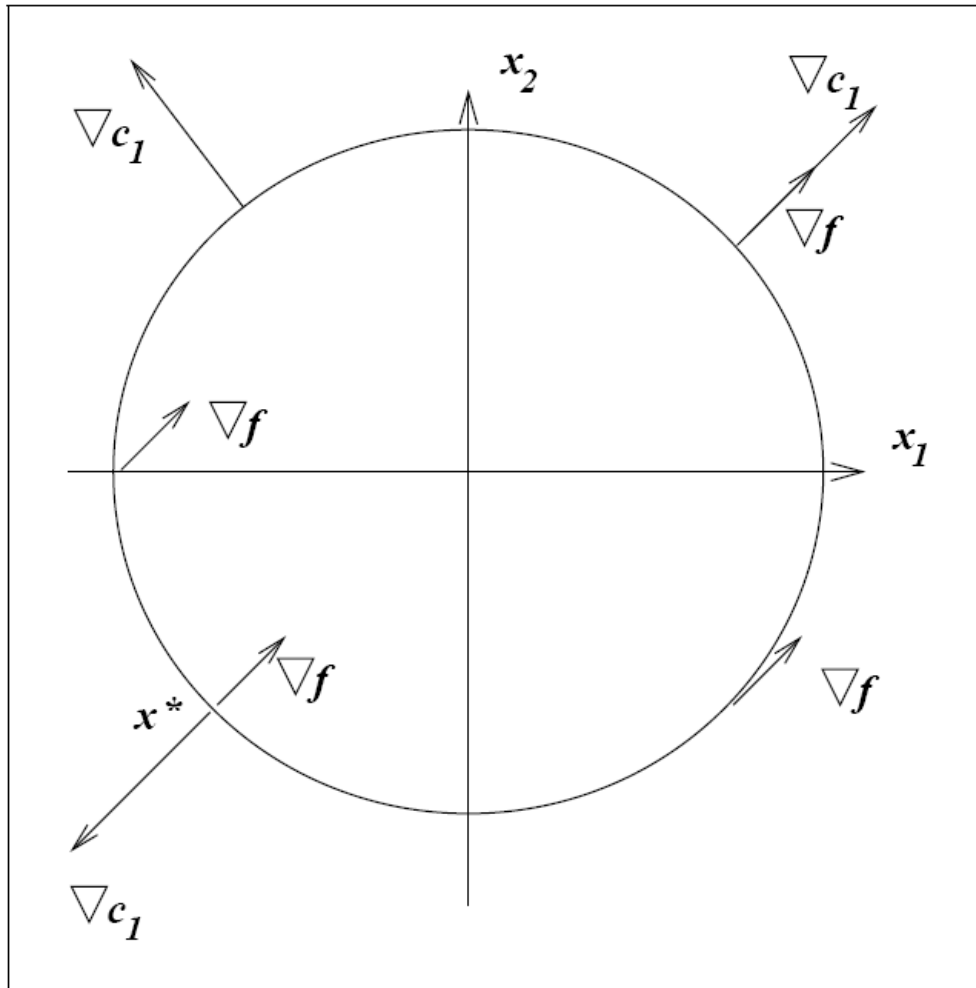
At a feasible point x , the inequality constraint $i \in I$ is said to be *active* if $c_i(x) = 0$ and *inactive* if the strict inequality $c_i(x) > 0$ is satisfied.

A SINGLE EQUALITY CONSTRAINT

$$\min x_1 + x_2 \quad \text{s.t.} \quad x_1^2 + x_2^2 - 2 = 0 \quad (1)$$


$$\begin{aligned} f(x) &= x_1 + x_2, \quad I = \emptyset, \quad E = \{1\} \\ c_1(x) &= x_1^2 + x_2^2 - 2 \end{aligned}$$

Example 1 (cont'd)



At the solution x^* , the constraint normal $\nabla c_1(x^*)$ is parallel to $\nabla f(x^*)$.

That is, there is a scalar λ_1^* such that

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*)$$

$$\text{Here, } \lambda_1^* = -\frac{1}{2}$$

Example 1 (cont'd)

To retain feasibility with respect to the function $c_1(x) = 0$, we require that $c_1(x + d) = 0$; that is,

$$0 = c_1(x + d) \approx c_1(x) + \nabla c_1(x)^T d = \nabla c_1(x)^T d$$

Similarly, a direction of improvement must produce a decrease in f , so that

$$0 > f(x + d) - f(x) \approx \nabla f(x)^T d$$

So to first order,
$$\begin{cases} \nabla c_1(x)^T d = 0 \\ \nabla f(x)^T d < 0 \end{cases}$$

if there exists a direction d that satisfies both, then?
--

Example 1 (cont'd)

Therefore, a *necessary condition* for optimality for the considered problem is:

there exist *no direction* d satisfying both $\nabla c_1(x)^T d = 0$
and $\nabla f(x)^T d < 0$.



In which case?

The only way that such a direction cannot exist is if $\nabla f(x)$ and $\nabla c_1(x)$ are *parallel*.



$\nabla f(x) = \lambda_1 \nabla c_1(x)$ holds at x , for some scalar λ_1 .

Example 1 (cont'd)

By introducing the *Lagrangian function*

$$L(x, \lambda_1) = f(x) - \lambda_1 c_1(x)$$



$$\nabla_x L(x, \lambda_1) = \nabla f(x) - \lambda_1 \nabla c_1(x)$$

Lagrange multiplier

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*) \iff \nabla_x L(x^*, \lambda_1^*) = 0$$

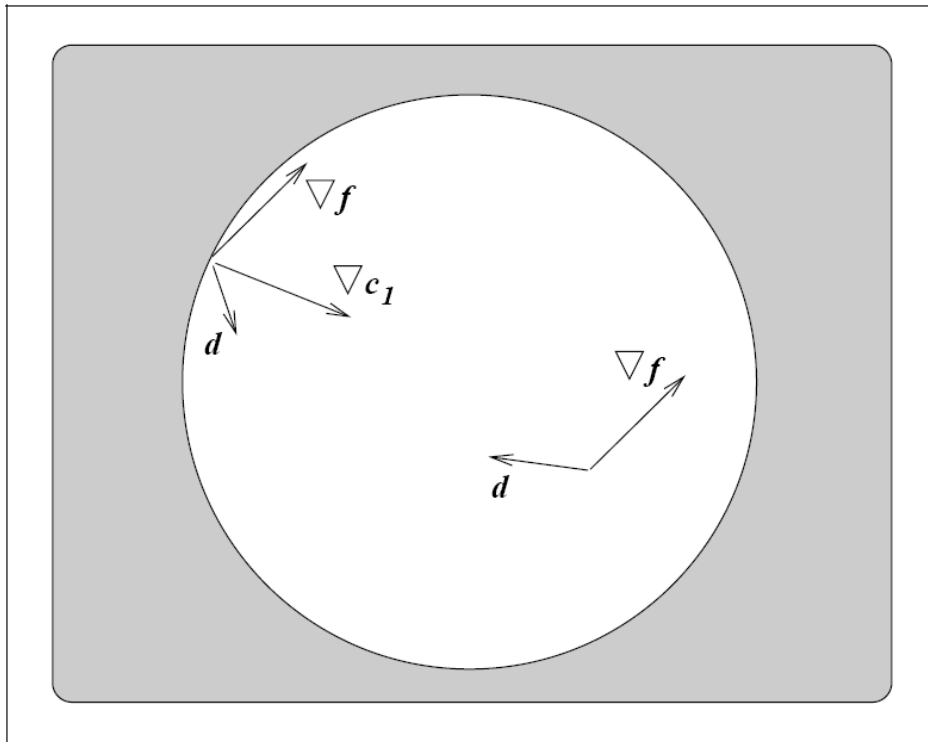
Is this sufficient to be necessary condition?

This observation suggests that we can search for solutions of the equality-constrained problem by **searching for stationary points of the Lagrangian function**.

Example 2

A SINGLE INEQUALITY CONSTRAINT

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0 \quad (2)$$



By inspection, we see that the solution is still $(-1, -1)$ and that the condition

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*)$$

holds for the value $\lambda_1^* = \frac{1}{2}$.

the sign is different!

Example 2 (cont'd)

As before, we conjecture that a given feasible point x is *not optimal* if we can find a step d that both retains feasibility and decreases the objective function f to first order.

- The direction d improves the objective function, to first order, if

$$\nabla f(x)^T d < 0$$

- The direction d retains feasibility, to first order, if

$$0 \leq c_1(x+d) \approx c_1(x) + \nabla c_1(x)^T d$$

We consider two cases in determining whether a direction d exists that satisfies the above two conditions.

Example 2 (cont'd)

Case I: in which x lies strictly inside the circle, so that the strict inequality $c_1(x) > 0$ holds.



As long as the length of the vector d is sufficiently small, it satisfies the feasible condition.

In particular,

$$d = -c_1(x) \frac{\nabla f(x)}{\|\nabla c_1(x)\| \|\nabla f(x)\|}$$

satisfies both whenever $\nabla f(x^*) \neq 0$.

The only situation in which such a direction fails to exist is when

$$\nabla f(x) = 0$$

Example 2 (cont'd)

Case II: in which x lies on the boundary of the circle, so that $c_1(x) = 0$. The discussed conditions therefore become

$$\nabla f(x)^T d < 0$$

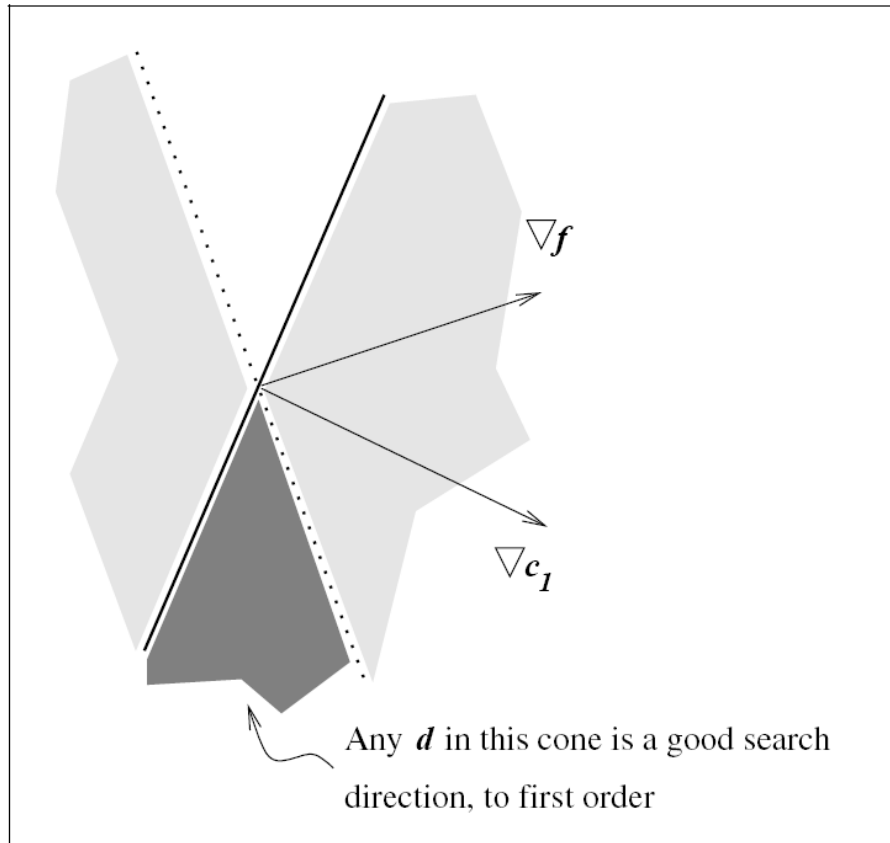
$$\nabla c_1(x)^T d \geq 0$$

The two regions fail to intersect only when

$$\nabla f(x) = \lambda_1 \nabla c_1(x)$$

for some $\lambda_1 \geq 0$

same direction



Example 2 (cont'd)

The optimality conditions for both cases I and II can be summarized neatly with reference to the Lagrangian function.

When no first-order feasible descent direction exists at some point x^* , we have that

$$\nabla_x L(x^*, \lambda_1^*) = 0, \quad \text{for some } \lambda_1^* \geq 0$$

where we also require that

$$\lambda_1^* c_1(x^*) = 0 \quad \left\{ \begin{array}{l} \text{Case I} \\ \text{Case II} \end{array} \right.$$

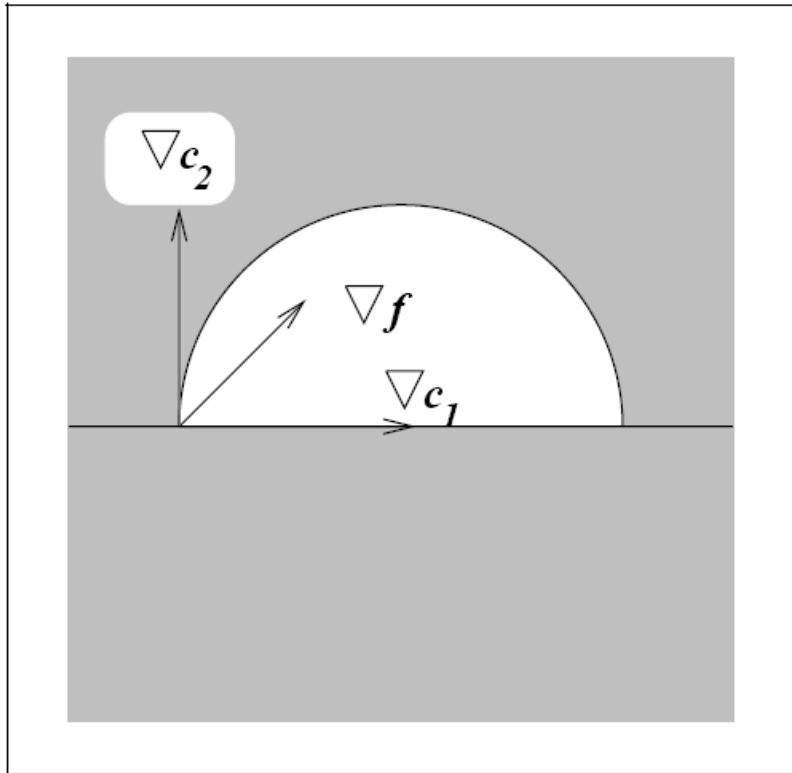


complementarity condition

Example 3

TWO INEQUALITY CONSTRAINTS

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0, \quad x_2 \geq 0 \quad (3)$$



It is easy to see that the solution lies at

$$\left(-\sqrt{2}, 0\right)^T$$

a point at which both two constraints are active.

Example 3 (cont'd)

A direction d is a *feasible descent* direction, to first order, if it satisfies the following conditions:

$$\nabla c_i(x)^T d \geq 0, \quad i \in I = \{1, 2\}, \quad \nabla f(x)^T d < 0$$

However, no such direction can exist when $x = (-\sqrt{2}, 0)^T$.

Next let us see how the Lagrangian and its derivatives behave for the concerned problem and its solution point.

$$L(x, \lambda) = f(x) - \lambda_1 c_1(x) - \lambda_2 c_2(x)$$

where $\lambda = (\lambda_1, \lambda_2)^T$ is the vector of Lagrange multipliers.

Example 3 (cont'd)

$$\nabla_x L(x^*, \lambda^*) = 0, \quad \text{for some } \lambda^* \geq 0$$

$$\lambda_1^* c_1(x^*) = 0, \quad \lambda_2^* c_2(x^*) = 0$$

When $x^* = (-\sqrt{2}, 0)^T$, we have

$$\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla c_1(x^*) = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix}, \quad \nabla c_2(x^*) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

so that it is easy to verify that $\nabla_x L(x^*, \lambda^*) = 0$ when we select

λ^* as follows:

$$\lambda^* = \begin{bmatrix} 1/(2\sqrt{2}) \\ 1 \end{bmatrix}$$

Example 3 (cont'd)

We consider now some other feasible points that are *not* solutions of (3).

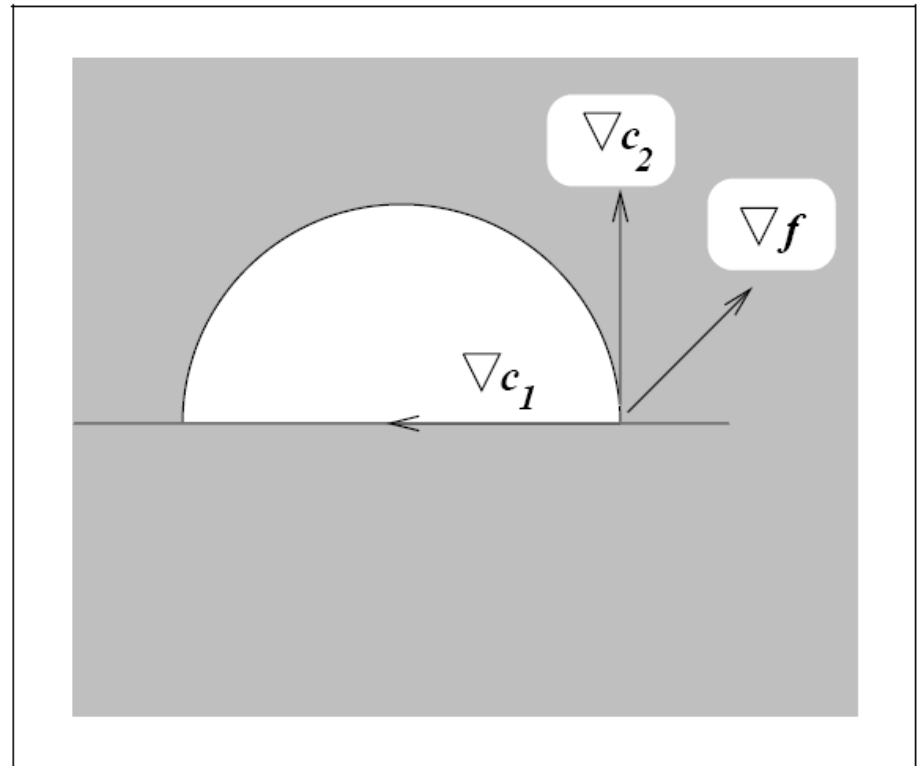
- $x = (\sqrt{2}, 0)^T$

again both constraints are active, but $\nabla f(x) \dots$

— One first-order feasible descent direction from this point is simply

$$d = (-1, 0)^T$$

there are many others.



Example 3 (cont'd)

- $x = (1, 0)^T$

at which only the second constraint c_2 is active.

To be a feasible descent direction, to first order, the direction d must satisfy:

$$1 + \nabla c_1(x)^T d \geq 0, \quad \nabla c_2(x)^T d \geq 0, \quad \nabla f(x)^T d < 0$$

$$\nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla c_2(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \Longrightarrow \quad d = \left(-\frac{1}{2}, \frac{1}{4} \right)$$

To show that optimality conditions fail, we note that

$$c_1(x) > 0 \quad \Rightarrow \quad \lambda_1 = 0 \quad \nabla f(x) - \lambda_2 \nabla c_2(x) = 0 \quad \text{exist?}$$

First-order necessary conditions

The three examples above suggest that a number of conditions are important in the characterization of solutions for the constrained optimization problem.

- the relation $\nabla_x L(x, \lambda) = 0$
- the nonnegativity of λ_i for all inequality constraints $c_i(x)$
- the complementarity condition $\lambda_i c_i(x) = 0$ that is required for all the inequality constraints.

