



EL3300/SF3849 Convex Optimization with Engineering Applications

Lecture 9: Smooth convex unconstrained and equality-constrained minimization

Anders Forsgren

Notes

Unconstrained convex program

Consider a convex optimization problem on the form

$$(CP) \quad \underset{x \in R^n}{\text{minimize}} \quad f(x),$$

where $f : R^n \rightarrow R$ is convex and twice continuously differentiable.

Proposition

Let p^* denote the optimal value of (CP). Let x^* in R^n , let $m = \eta_{\min}(\nabla^2 f(x))$ and let $M = \eta_{\max}(\nabla^2 f(x))$. Then,

$$f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \leq p^* \leq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2.$$

Note that $\tilde{y} = x - (1/m)\nabla f(x)$ minimizes $\nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|_2^2$.

The direction $-(1/m)\nabla f(x)$ is a **steepest descent** step.

Notes

Iterative methods

$$(CP) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in R^n, \end{array}$$

where $f \in C^2$, f convex on R^n .

An iterative method generates x_0, x_1, x_2, \dots such that $\lim_{k \rightarrow \infty} x_k = x^*$, where $\nabla f(x^*) = 0$.

Terminates when suitable convergence criteria is fulfilled, e.g., $\|\nabla f(x_k)\| < \epsilon$.

Notes

Linesearch methods

A **linesearch method** generates in each iteration a **search direction** and performs a **linesearch** along the search direction.

Iteration k takes the following form at x_k .

- 1 Compute search direction p_k such that $\nabla f(x_k)^T p_k < 0$.
- 2 Approximately solve $\min_{\alpha \geq 0} f(x_k + \alpha p_k)$, which gives α_k .
- 3 $x_{k+1} \leftarrow x_k + \alpha_k p_k$.

Different methods vary in choice of p_k and α_k .

Notes

Classes of line search methods

We will initially consider two fundamental methods.

- The **steepest-descent method**, where $p_k = -\nabla f(x_k)$, and
- **Newton's method**, where $\nabla^2 f(x_k)p_k = -\nabla f(x_k)$.

Steepest descent: + Search direction inexpensive to compute,
- Slow convergence.

Newton's method: - Search direction more expensive to compute,
+ Faster convergence.

There are methods "in-between", e.g., **quasi-Newton methods** that aim at mimicking Newton's method without computing second derivatives.

Notes

Quadratic objective function

Consider model problem with quadratic objective function

$$(QP) \quad \begin{array}{ll} \text{minimize} & f(x) = \frac{1}{2}x^T Hx + c^T x \\ \text{subject to} & x \in \mathbb{R}^n, \end{array}$$

where $H \succeq 0$.

Proposition

The following holds for (QP) depending on H and c :

- If $H \succ 0$ then (QP) has a unique minimizer x^* given by $Hx^* = -c$.
- If $H \succeq 0$, $H \not\succeq 0$ each x^* that fulfills $Hx^* = -c$ is a global minimizer to (QP). (There may possibly be no such x^*).

Proof.

The condition $\nabla f(x^*) = 0$ gives the results. \square

We assume $H \succ 0$ in the discussion.

Notes

Linesearch method on quadratic objective function

Consider (QP) with $f(x) = \frac{1}{2}x^T Hx + c^T x$, where $H \succ 0$. We obtain:

x^* minimizer to (QP) $\iff 0 = \nabla f(x^*) = Hx^* + c$.

Suppose search direction p_k satisfies $\nabla f(x_k)^T p_k < 0$.

Let $\varphi(\alpha) = f(x_k + \alpha p_k) = f(x_k) + \alpha \nabla f(x_k)^T p_k + \frac{\alpha^2}{2} p_k^T H p_k$.

Then $\alpha_k = -\frac{\nabla f(x_k)^T p_k}{p_k^T H p_k}$ gives the minimizer to

$\min_{\alpha \geq 0} f(x_k + \alpha p_k)$, i.e., we can perform **exact linesearch**.

Notes

Steepest descent on quadratic objective function

Assume that $f(x) = \frac{1}{2}x^T Hx + c^T x$, where $H \succ 0$.

Further assume that steepest descent with exact linesearch is

used, i.e., $p_k = -\nabla f(x_k)$ and $\alpha_k = -\frac{\nabla f(x_k)^T p_k}{p_k^T H p_k}$.

Then it can be shown that

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\text{cond}(H) - 1}{\text{cond}(H) + 1} \right)^2 (f(x_k) - f(x^*)).$$

$\text{cond}(H) \gg 1 \Rightarrow \frac{\text{cond}(H) - 1}{\text{cond}(H) + 1} \approx 1$, i.e., slow **linear convergence**.

For a nonlinear function, we typically get slow linear convergence, where H is replaced by $\nabla^2 f(x_k)$.

Notes

Speed of convergence

Definition

Assume that $x_k \in \mathbb{R}^n$, $k = 0, 1, \dots$, and assume that $\lim_{k \rightarrow \infty} x_k = x^*$. We say that $\{x_k\}_{k=0}^{\infty}$ converges to x^* with speed of convergence r if

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^r} = C, \quad \text{where } C < \infty.$$

We have $\|x_{k+1} - x^*\| \approx C \cdot \|x_k - x^*\|^r$.

We want r large (and C close to zero). Of interest:

- $r = 1$, $0 < C < 1$, linear convergence. (Steepest descent.)
- $r = 1$, $C = 0$, superlinear convergence. (Quasi-Newton.)
- $r = 2$, quadratic convergence. (Newton's method.)

Notes

Newton's method for solving a nonlinear equation

Consider solving the nonlinear equation $\nabla f(u) = 0$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f \in \mathcal{C}^2$.

Then, $\nabla f(u + p) = \nabla f(u) + \nabla^2 f(u)p + o(\|p\|)$.

Linearization given by $\nabla f(u) + \nabla^2 f(u)p$.

Choose p so that $\nabla f(u) + \nabla^2 f(u)p = 0$, i.e., solve $\nabla^2 f(u)p = -\nabla f(u)$.

A Newton iteration takes the following form for a given u .

- 1 p solves $\nabla^2 f(u)p = -\nabla f(u)$.
- 2 $u \leftarrow u + p$.

(The nonlinear equation need not be a gradient.)

Notes

Speed of convergence for Newton's method

Theorem

Assume that $f \in C^3$ and that $\nabla^2 f(u^*)$ is nonsingular. Then, if Newton's method (with steplength one) is started at a point sufficiently close to u^* , then it is well defined and converges to u^* with convergence rate at least two, i.e., there is a constant C such that $\|u_{k+1} - u^*\| \leq C\|u_k - u^*\|^2$.

The proof can be given by studying a Taylor-series expansion,

$$\begin{aligned} u_{k+1} - u^* &= u_k - \nabla^2 f(u_k)^{-1} \nabla f(u_k) - u^* \\ &= -\nabla^2 f(u_k)^{-1} (\nabla f(u^*) - \nabla f(u_k) + \nabla^2 f(u_k)(u^* - u_k)). \end{aligned}$$

For u_k sufficiently close to u^* ,

$$\|\nabla f(u^*) - \nabla f(u_k) + \nabla^2 f(u_k)(u^* - u_k)\| \leq \tilde{C}\|u_k - u^*\|^2.$$

Notes

One-dimensional example for Newton's method

For a positive number d , consider computing $1/d$ by minimizing $f(u) = du - \ln u$.

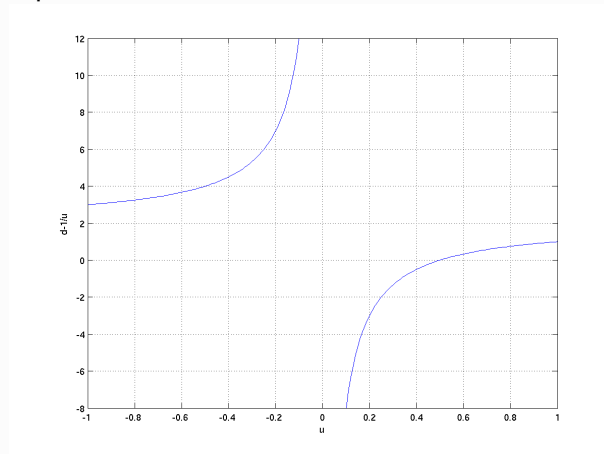
Then, $f'(u) = d - \frac{1}{u}$, $f''(u) = \frac{1}{u^2}$. We see that $u^* = \frac{1}{d}$.

$$u_{k+1} = u_k - \frac{f'(u_k)}{f''(u_k)} = u_k - \frac{d - \frac{1}{u_k}}{\frac{1}{u_k^2}} = 2u_k - u_k^2 d.$$

$$\text{Then, } u_{k+1} - \frac{1}{d} = 2u_k - u_k^2 d - \frac{1}{d} = -d \left(u_k - \frac{1}{d} \right)^2.$$

Notes

Graphical picture for $d = 2$.



Notes

Sufficient descent direction

For a direction p_k to ensure convergence is must be a **sufficient descent direction**. Typical conditions are

$$-\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\| \|p_k\|} \geq \sigma, \quad \text{where } \sigma \text{ is a positive constant .}$$

This means that p_k must be “sufficiently similar” to the negative gradient.

For search direction p_k from $B_k p_k = -\nabla f(x_k)$ this is required by ensuring that $\|B_k\| \leq M$ and $\|B_k^{-1}\| \leq m$, where m and M are positive constants.

In a modified Newton method modifications of $\nabla^2 f(x_k)$ can be made, if needed, by a modified Cholesky factorization.

Notes

Linesearch

In the linesearch α_k is determined as an approximate solution to $\min_{\alpha \geq 0} \varphi(\alpha)$, where $\varphi(\alpha) = f(x_k + \alpha p_k)$. We want $f(x_{k+1}) < f(x_k)$, i.e., $\varphi(\alpha_k) < \varphi(0)$. This is not sufficient to ensure convergence.

Example requirement for step not too long:

$$\begin{aligned} \varphi(\alpha) &\leq \varphi(0) + \mu\alpha\varphi'(0), \text{ i.e.,} && \text{(Armijo condition)} \\ f(x_k + \alpha p_k) &\leq f(x_k) + \mu\alpha \nabla f(x_k)^T p_k, \\ \text{where } \mu &\in (0, \frac{1}{2}). \end{aligned}$$

Example requirement for step not too short:

$$\begin{aligned} |\varphi'(\alpha)| &\leq -\eta\varphi'(0), \text{ i.e.,} && \text{(Wolfe condition)} \\ |\nabla f(x_k + \alpha p_k)^T p_k| &\leq -\eta \nabla f(x_k)^T p_k, \end{aligned}$$

where $\eta \in (\mu, 1)$. Alternative requirement for not too short step:

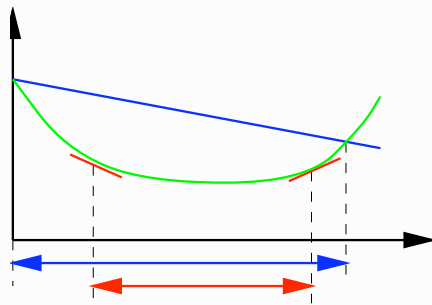
Take smallest nonnegative integer i such that

$$f(x_k + 2^{-i} p_k) \leq f(x_k) + \mu 2^{-i} \nabla f(x_k)^T p_k. \quad \text{("Backtracking")}$$

Notes

Illustration of linesearch conditions

The linesearch conditions of Wolfe-Armijo type can be illustrated in the following picture.



Notes

Line search conditions

To find $\bar{\alpha}$ such that the Wolfe and Armijo conditions are fulfilled we may consider $\hat{\varphi}(\alpha) = \varphi(\alpha) - \varphi(0) - \mu\alpha\varphi'(0)$.

Then $\hat{\varphi}(0) = 0$ and $\hat{\varphi}'(0) < 0$. In addition, there must exist $\bar{\alpha} > 0$ such that $\hat{\varphi}(\bar{\alpha}) = 0$, otherwise φ is unbounded from below.

By the mean-value theorem there is an $\hat{\alpha} \in (0, \bar{\alpha})$ such that $\hat{\varphi}(\hat{\alpha}) < 0$ and $\hat{\varphi}'(\hat{\alpha}) = 0$.

Since $\mu < \eta$ we obtain $\varphi(\alpha) \leq \varphi(0) + \mu\alpha\varphi'(0)$ and $|\varphi'(\alpha)| \leq -\eta\varphi'(0)$ for α in a neighborhood of $\hat{\alpha}$.

For example, bisection in combination with polynomial interpolation can be used on $\hat{\varphi}$ to find a suitable α .

Notes

Newton's method and steepest descent

The steepest descent direction solves

$$\begin{array}{ll} \text{minimize} & \nabla f(x)^T p \\ \text{subject to} & p^T p \leq 1. \end{array}$$

The Newton direction solves

$$\begin{array}{ll} \text{minimize} & \nabla f(x)^T p \\ \text{subject to} & p^T \nabla^2 f(x) p \leq 1. \end{array}$$

The Newton step is a steepest-descent step in the norm defined by $\nabla^2 f(x)$, i.e.,

$$\|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}.$$

Notes

Self-concordant functions

When proving polynomial complexity of **interior methods** for convex optimization, the notion of **self-concordant functions** is an important concept.

Definition

A three times differentiable function $f : C \rightarrow R$, which is convex on the convex set C , is **self-concordant** if $|f'''(x)| \leq 2f''(x)^{3/2}$.

In essence, this means that the third derivatives are not “too large”.

An important self-concordant function is $f(x) = -\ln x$ for $x > 0$.

Notes

Suggested reading

Suggested reading in the textbook:

- Sections 9.1–9.7.

Notes

Equality-constrained convex program

Consider a convex optimization problem on the form

$$(CP_{=}) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and twice continuously differentiable.

If the Lagrangian function is defined as $l(x, \lambda) = f(x) - \lambda^T(Ax - b)$, the first-order optimality conditions are $\nabla l(x, \lambda) = 0$. We write them as

$$\begin{pmatrix} \nabla_x l(x, \lambda) \\ -\nabla_\lambda l(x, \lambda) \end{pmatrix} = \begin{pmatrix} \nabla f(x) - A(x)^T \lambda \\ Ax - b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Notes

Newton iteration

A Newton iteration on the optimality conditions takes the form

$$\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\nu \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - A^T \lambda \\ Ax - b \end{pmatrix}.$$

We may use

$$\left\| \begin{pmatrix} \nabla f(x) - A^T \lambda \\ Ax - b \end{pmatrix} \right\|^2$$

as **merit function**, i.e., to measure how “good” a point is.

Notes

Variable elimination

Note that for a feasible point \bar{x} , it holds that $A(x - \bar{x}) = 0$ for all feasible x . Let Z be a matrix whose columns form a basis for $\text{null}(A)$. Then $x = \bar{x} + Zv$, with a one-to-one correspondence between x and v .

Let $\varphi(v) = f(\bar{x} + Zv)$. We may then rewrite the problem as

$$(CP'_{=}) \quad \underset{v \in \mathbb{R}^{n-m}}{\text{minimize}} \quad \varphi(v).$$

Differentiation gives $\nabla\varphi(v) = Z^T \nabla f(\bar{x} + Zv)$,
 $\nabla^2\varphi(v) = Z^T \nabla^2 f(\bar{x} + Zv) Z$.

This is an unconstrained problem. We may solve $(CP'_{=})$ and identify $x^* = \bar{x} + Zv^*$, where v^* is associated with $(CP'_{=})$.

$Z^T \nabla f(x)$ is called the **reduced gradient** to f in x .

$Z^T \nabla^2 f(x) Z$ is called the **reduced Hessian** to f in x .

Notes

First-order optimality conditions as a system of equations, cont.

The resulting Newton system may equivalently be written as

$$\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -(\lambda + \nu) \end{pmatrix} = \begin{pmatrix} -\nabla f(x) \\ -(Ax - b) \end{pmatrix},$$

alternatively

$$\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\lambda^+ \end{pmatrix} = \begin{pmatrix} -\nabla f(x) \\ -(Ax - b) \end{pmatrix}.$$

We prefer the form with λ^+ , since it can be directly generalized to problems with inequality constraints.

Notes

Quadratic programming with equality constraints

Compare with an equality-constrained quadratic programming problem

$$(EQP) \quad \begin{array}{ll} \text{minimize} & \frac{1}{2}p^T H p + c^T p \\ \text{subject to} & A p = b, \\ & p \in \mathbb{R}^n, \end{array}$$

where the unique optimal solution p and multiplier vector λ^+ are given by

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\lambda^+ \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix},$$

if $Z^T H Z \succ 0$ and A has full row rank, where Z is a matrix whose columns form a basis for $\text{null}(A)$.

Notes

Newton iteration and equality-constrained quadratic program

$$\text{Compare} \quad \begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\lambda^+ \end{pmatrix} = \begin{pmatrix} -\nabla f(x) \\ -(Ax - b) \end{pmatrix}$$

$$\text{with} \quad \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\lambda^+ \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}.$$

$$\text{Identify:} \quad \begin{array}{lll} \nabla^2 f(x) & \longleftrightarrow & H \\ \nabla f(x) & \longleftrightarrow & c \\ A & \longleftrightarrow & A \\ -(Ax - b) & \longleftrightarrow & b. \end{array}$$

Notes

Newton iteration as a QP problem

A Newton iteration for solving the first-order necessary optimality conditions to $(CP_{=})$ may be viewed as solving the QP problem

$$(QP_{=}) \quad \begin{array}{ll} \text{minimize} & \frac{1}{2}p^T \nabla^2 f(x)p + \nabla f(x)^T p \\ \text{subject to} & Ap = -(Ax - b), \\ & p \in R^n, \end{array}$$

and letting $x^+ = x + p$, and λ^+ are given by the multipliers of $(QP_{=})$.

Problem $(QP_{=})$ is well defined with unique optimal solution p and multiplier vector λ^+ if $Z^T \nabla^2 f(x)Z \succ 0$ and A has full row rank, where Z is a matrix whose columns form a basis for $\text{null}(A)$.

Notes

An SQP iteration for problems with equality constraints

Given x, λ such that $Z^T \nabla^2 f(x)Z \succ 0$ and A has full row rank, a Newton iteration takes the following form.

- 1 Compute optimal solution p and multiplier vector λ^+ to

$$(QP_{=}) \quad \begin{array}{ll} \text{minimize} & \frac{1}{2}p^T \nabla^2 f(x)p + \nabla f(x)^T p \\ \text{subject to} & Ap = -(Ax - b), \\ & p \in R^n, \end{array}$$

- 2 $x \leftarrow x + p, \quad \lambda \leftarrow \lambda^+$.

We call this method **sequential quadratic programming** (SQP).

NB! $(QP_{=})$ is solved by solving a system of linear equations.

NB! x and λ have given numerical values in $(QP_{=})$.

Notes

Speed of convergence for SQP method for equality-constrained problems

Theorem

Assume that $f \in C^3$ is convex on R^n and that $A \in R^{m \times n}$ has full row rank. Further, assume that x^* is a minimizer of $(CP_=)$ such that $Z^T \nabla^2 f(x) Z \succeq 0$, where Z is a matrix whose columns form a basis for $\text{null}(A)$. If the SQP method (with steplength one) is started sufficiently close to x^* , λ^* , then it is well defined and converges to x^* , λ^* with convergence rate at least two.

Proof.

In a neighborhood of x^* , λ^* it holds that $Z^T \nabla^2 f(x) Z \succ 0$ and $\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix}$ is nonsingular. The subproblem $(QP_=)$ is hence well defined and the result follows from the quadratic rate of convergence of Newton's method. \square

Notes

Nonlinearly constrained convex program

Consider a convex optimization problem on the form

$$(CP) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \geq 0, \quad i \in \mathcal{I}, \quad \mathcal{I} \cup \mathcal{E} = \{1, \dots, m\}, \\ & a_i^T x - b_i = 0, \quad i \in \mathcal{E}, \quad \mathcal{I} \cap \mathcal{E} = \emptyset, \\ & x \in R^n, \end{array}$$

where $f : R^n \rightarrow R$ and $-g_i : R^n \rightarrow R$ are convex and twice continuously differentiable on R^n .

The inequality constraints give an added combinatorial problem of identifying the constraints that are active at the solution.

One way of dealing with inequality constraints is via a **barrier transformation**.

Notes

Barrier transformation

Consider replacing an inequality constraint $g_i(x) \geq 0$ by a **logarithmic barrier term** $-\ln(g_i(x))$ added to the objective function.

Other barrier terms are possible. The logarithmic barrier term is the "canonic" choice.

The effect is a perturbed problem where an infinite cost is incurred as $g_i(x) \rightarrow 0$.

The weight which we put on the barrier term is denoted by μ and referred to as the **barrier parameter**.

The combinatorial effect is removed at the expense of a perturbation of the original problem.

Notes

The problem resulting from a barrier transformation

For a positive barrier parameter μ , the barrier transformed problem becomes

$$(CP_\mu) \quad \begin{array}{ll} \underset{x \in R^n}{\text{minimize}} & f(x) - \mu \sum_{i \in \mathcal{I}} \ln(g_i(x)) \\ & a_i^T x - b_i = 0, \quad i \in \mathcal{E}, \end{array} \quad \begin{array}{l} \mathcal{I} \cup \mathcal{E} = \{1, \dots, m\}, \\ \mathcal{I} \cap \mathcal{E} = \emptyset. \end{array}$$

Note that convexity is preserved.

Proposition

Let $g_i : R^n \rightarrow R$ be a concave function on R^n . Then, $-\ln(g_i(x))$ is a convex function on the convex set $\{x \in R^n : g_i(x) > 0\}$.

Notes

The barrier trajectory

Under suitable assumptions, the barrier transformed problem has a unique optimal solution $x(\mu)$ and corresponding Lagrange multipliers $\lambda_i(\mu)$, $i \in \mathcal{E}$, for each $\mu > 0$.

In this situation, the **barrier trajectory** is defined as the set $\{x(\mu) : \mu > 0\}$. The barrier trajectory is sometimes referred to as the **central path**.

Theorem

Under suitable assumptions, the barrier trajectory is well defined and it holds that $\lim_{\mu \rightarrow 0} x(\mu) = x^$, $\lim_{\mu \rightarrow 0} \mu/g_i(x(\mu)) = \lambda_i^*$, $i \in \mathcal{I}$, and $\lim_{\mu \rightarrow 0} \lambda_i(\mu) = \lambda_i^*$, $i \in \mathcal{E}$, where x^* is an optimal solution to (CP), and λ^* is an associated Lagrange multiplier vector.*

Hence, the barrier trajectory converges to an optimal solution.

Notes

A primal approach: Sequential unconstrained minimization

We may now apply the methods outlined previously for (approximately) solving a sequence of unconstrained minimization problems for decreasing values of the barrier parameter μ .

This is sometimes referred to as sequential unconstrained minimization techniques. We shall refer to this approach as a primal approach. The method is an **interior method**, i.e., it generates points that lie in the (relative) interior of the feasible set.

Let $f_\mu(x) = f(x) - \mu \sum_{i \in \mathcal{I}} \ln(g_i(x))$. Then,

$$\nabla^2 f_\mu(x) = \nabla^2 f(x) - \sum_{i \in \mathcal{I}} \frac{\mu}{g_i(x)} \nabla^2 g_i(x) + \sum_{i \in \mathcal{I}} \frac{\mu}{(g_i(x))^2} \nabla g_i(x) \nabla g_i(x)^T.$$

As $\mu \rightarrow 0$, $\nabla^2 f_\mu(x(\mu))$ becomes increasingly ill-conditioned in general, with the condition number tending to infinity.

Notes
